Essays in Financial Econometrics and Forecasting

Ekaterina Smetanina

Faculty of Economics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Queens' College

October 2018

To my family

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration, except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution. It does not exceed the prescribed word limit of 60,000 words.

Ekaterina Smetanina

October 2018

Acknowledgements

First and foremost I would like to thank my supervisor, Oliver Linton, for taking me as a student and for his constant encouragement, guidance and support throughout my time in Cambridge. I am deeply grateful to my advisor, Alexey Onatskiy, for always finding time to discuss my work and for his great advice, especially during the job market. I feel very privileged to have had them both as my advisors.

I would especially like to thank Anders Rahbek for taking a great interest in my work and for inviting me to Copenhagen over the years to discuss and present my work. I am also thankful to Anders for encouraging me to work on what became the second chapter of this dissertation. I would also like to thank Rasmus Søndergaard Pedersen for our many discussions during my visits to Copenhagen and his many comments and suggestions on my work. My many conversations with Anders and Rasmus in Cambridge and Copenhagen together with their encouragement and invaluable guidance were essential in writing the second chapter of this dissertation.

Very special thanks also goes to Tim Bollerslev who kindly agreed to my visit to Duke University during my PhD studies. He made my Duke visit an exciting and intellectually enriching time. I would like to thank him for our many interesting conversations we had at Duke and especially for his support during my job market. He has inspired me and my work greatly.

I thank Andrew Patton, who always took time to comment on my work. Andrew has been always very generous with his time during my visit at Duke, and I especially thank him for our many discussions about forecasting and for his constant encouragement of my ideas.

I would like to thank Dennis Kristensen and Alexey Onatskiy for agreeing to be on my committee and for their time to comment on my work.

I would further like to thank the Faculty of Economics and my colleagues in DAE for sharing many working days, countless coffee breaks and innumerable interesting discussions we had over the years. I would particularly like to thank Anil Ari, Jeroen Dalderop, Axel Gottfries, Simon Lloyd, Sam Mann, Rafe Martyn, David Minarsch, Monica Petrescu, Mikhail Safronov, Peng Zhang and Ruochen Wu. I am extremely grateful to River Chen for being a great friend not only to me but to all my family.

I would also like to acknowledge the financial support during my PhD studies from the

Cambridge INET Institute and the Royal Economic Society.

Most of all, I am indebted to my family who are always a constant source of support and enormous love. My mother has been an inspiration throughout my life, and although even a lifetime will not be enough to come close to her achievements, I hope this small achievement of mine will make her a little bit more proud. Last but not least, I would like to thank Jason for his continuous support.

Abstract

This dissertation deals with issues of forecasting in financial markets. The first part of my dissertation is motivated by the observation that most parametric volatility models follow Engle's (1982) original idea of modelling the volatility of asset returns as a function of only past information. However, current returns are potentially quite informative for forecasting, yet are excluded from these models. The first and second chapters of this dissertation try to address this question from both a theoretical and an empirical perspective. The second part of this dissertation deals with the important issue of forecast evaluation and selection in unstable environments, where it is known that the existing methodology can generate spurious and potentially misleading results. In my third chapter, I develop a new methodology for forecast evaluation and selection in such an environment.

In the first chapter, *Real-time GARCH*, I propose a new parametric volatility model, which retains the simple structure of GARCH models, but models the volatility process as a mixture of past and current information as in the spirit of Stochastic Volatility (SV) models. This provides therefore a link between GARCH and SV models. I show that with this new model I am able to obtain better volatility forecasts than the standard GARCH-type models; improve the empirical fit of the data, especially in the tails of the distribution; and make the model faster in its adjustment to the new unconditional level of volatility. Further, the new model offers a much needed framework for specification testing as it nests the standard GARCH models. This chapter has been published in the *Journal of Financial Econometrics* (Smetanina E., 2017, Real-time GARCH, *Journal of Financial Econometrics*, 15(4), 561-601.)

In chapter 2, *Asymptotic Inference for Real-time GARCH(1,1) model*, I investigate the asymptotic properties of the Gaussian Quasi-Maximum-Likelihood estimator (QMLE) for the Real-time GARCH(1,1) model, developed in the first chapter of this dissertation. I establish the ergodicity and β-mixing properties of the joint process for squared returns and the volatility process. I also prove strong consistency and asymptotic normality for the parameter vector at the usual $\sqrt{T}$ rate. Finally, I demonstrate how the developed theory can be viewed as a generalisation of the QMLE theory for the standard GARCH(1,1) model.

In chapter 3, *Forecast Evaluation Tests in Unstable Environments*, I develop a new methodology for forecast evaluation and selection in the situations where the relative performance between models changes over time in an unknown fashion. Out-of-sample tests are widely used for evaluating models forecasts in economics and finance. Underlying these tests is often the assumption of constant relative performance between competing models, however this is invalid for many practical applications. In a world of changing relative performance, previous methodologies give rise to spurious and potentially misleading results, an example of which is the well-known "splitting point problem". I propose a new two-step methodology designed specifically for forecast

evaluation in a world of changing relative performance. In the first step I estimate the time-varying mean and variance of the series for forecast loss differences, and in the second step I use these estimates to construct new rankings for models in a changing world. I show that the new tests have high power against a variety of fixed and local alternatives.

Contents

Chapter 1

Real-time GARCH

1.1 Introduction

Volatility is widely used as a proxy for the risk associated with a financial asset, see e.g. French et al. (1987). Reliable estimation and forecasting of volatility is therefore crucial for many financial activities, such as risk management, portfolio choice and asset pricing. There are several main approaches to modelling the volatility of discrete financial time series: GARCH models (Engle, 1982; Bollerslev, 1986; Ding et al, 1993; Hansen et al., 2012, among others), Stochastic Volatility (SV) models (see Shephard, 2008 for overview), and hybrid models, e.g. Meddahi and Renault (2004). A main conceptual difference between the above approaches stems from the information structure they incorporate. Univariate GARCH models assume that the volatility of asset returns, σ_t, is a function of past information only, i.e. σ_t is $\mathcal{F}_{t-1}-$measurable, where $\mathcal{F}_{t-1}$ is the sigma-algebra induced by the history of returns up to time $t-1$. SV models assume that σ_t is $\mathcal{G}_t-$measurable, where $\mathcal{G}_t$ is the sigma-algebra induced by the history of returns as well as by the history of unobserved random shocks up to time t. The difference in the incorporated information structure is also in their nature: while GARCH models incorporate only past internal information (i.e. information generated only within the model itself) and are therefore deterministic, SV models generate a stochastic volatility process by allowing for external information in the form of unobserved random shocks that are independent from the shocks governing the returns process. As a result, SV models can be more flexible in fitting the data, however this comes at the cost of higher complexity involved in their estimation and inference. Contrasting with SV models, GARCH models are observation-driven. Hence they come with the advantage of having available many estimation methods, Quasi-Maximum Likelihood (QML) being the most popular, which accounts for their wider use among practitioners.

First remarked upon by Politis (2007), by not using all available internal information, in particular the current return, GARCH models make an inefficient use of information when forecasting the volatility of returns. An important implication of this is that GARCH models are poorly suited for situations of rapid changes in financial markets, for example when volatility changes rapidly to a new level, see e.g. Andersen et al. (2003), and Hansen et al. (2012). Until now it was assumed that all volatility models can be classified as either parameter-driven or observation-driven (see e.g. Cox, 1981 and Shephard, 1996), with a clear separation between the two. Since most GARCH models are observation-driven, it comes with a necessary condition that the process is modelled strictly in terms of the past observed information. Hence this limitation of GARCH models was believed to be inherent and unavoidable.

In this chapter I show that it is possible to efficiently utilise all available internal information in GARCH models, in particular incorporating the current return. I demonstrate that by doing so, I (i) can account for rapid changes in the unconditional level of volatility as the conditional distribution of returns has a time-varying kurtosis; (ii) outperform standard GARCH models in terms of both short-run (1 and 5 days ahead) and long-run (10 and 15 days ahead) out-of-sample volatility forecasts; (iii) provide a better empirical fit to the data, especially in the tails of the distribution; (iv) provide a conceptual link between SV and GARCH models; and (v) offer a much needed framework for specification testing of the standard GARCH models, which are nested in my framework.

To put things into context, consider the following model

$$r_t = \epsilon_t \lambda_t, \qquad \lambda_t \text{ is } \mathcal{F}_t - \text{measurable}, \tag{1.1}$$

where r_t is the (demeaned) return series, ϵ_t are i.i.d. random variables such that $E[\epsilon_t] = 0$, $E\left[\epsilon_t^2\right] = 1$, and $\mathcal{F}_t$ is the information set available at time t. Here I model volatility as a mixture of past as well as current information, i.e. λ_t is $\mathcal{F}_t-$measurable, where $\mathcal{F}_t$ contains only internal information. Compared to GARCH models, I use all information up to time t instead of time $t-1$. Compared to SV models, $\mathcal{F}_t$ contains only one source of randomness shared by the returns and volatility processes, which will allow me to retain a Quasi-Maximum Likelihood (QML) framework. The new model therefore can be thought of as a link between GARCH and SV models, as it nests the GARCH model as its special case, yet models the volatility process in the spirit of SV models where the two sources of randomness are perfectly correlated. While the new model combines the advantages of both GARCH and SV models in a unified framework, it is not strictly a GARCH nor a SV model, but rather it is in a new class of its own. I call this new model the "Real-time GARCH" model (RT-GARCH for short), indicating the fact that the most "current" information is contributing to the volatility process.

An important advantage of this framework is that it allows the shape of the conditional distribution of returns to be time-varying. This has two main implications. Firstly, unlike GARCH models where the conditional kurtosis of the error terms simply translates into the kurtosis of the returns, my model's conditional kurtosis is time-varying. Secondly, the conditional density of returns is no longer a scaled normal density even when the error term has a Gaussian density. The new density function has an extra shape parameter which determines the "peakedness", and/or thickness of the tails, of the returns distribution. This allows the new model to better capture tail behaviour of the returns. This shall play an important role for the precision of the out-of-sample Value-at-risk (VaR) and short- and long-run volatility forecasts.

Politis (2007) makes the first investigation of the implications of information loss for forecasting volatility. He develops a novel model-free normalising and variance stabilising (NoVaS) transformation of the initial time series of returns, by incorporating the current squared returns into the conditional variance process in order to improve volatility forecasts. Being a model-free specification, parameter estimates and statistical properties are not available. Thus direct comparison of the theoretical implications of this specification with existing discrete-time volatility models is not possible, and the important question of whether including current information in a more structured model would provide any improvements over the standard GARCH models was not addressed. I answer this question by studying the statistical properties and the empirical performance of the RT-GARCH model. I first show that it is possible to incorporate current information into GARCH-type models while retaining interpretation, and a good description of the key characteristics of financial data. I show that the new information, i.e. the current realisation of the current return (or some function of thereof), can be viewed in two ways: as a change in the information set, and as providing the conditional density of returns with an extra shape parameter, making it therefore time-varying.

In the empirical study, I estimate the new model on three datasets: IBM, GE and S&P 500 daily returns which span from the 2nd January 1998 (28th January 2003 for S&P500) to 1st December 2016. I find that accounting for current information in the volatility process plays an important role along several dimensions. Firstly, the RT-GARCH model outperforms standard GARCH-type models in terms of producing better short-run (1 and 5 day ahead) and especially long-run (10 and 15 days ahead) out-of-sample volatility forecasts. In particular, I compare 1-, 5-, 10- and 15-step ahead volatility forecasts with those of the GARCH(1,1) and GARCH(1,2) with standard normal and Student-t errors, APARCH(2,2) with Student-t errors, as well as Simple and Exponential NoVaS methodologies of Politis (2007). To evaluate the competing forecasts, I perform Hansen's (2011) Model Confidence Set (MCS) test and provide evidence that the RT-GARCH models always lie in the MCS for all horizons, while standard GARCH models are only occasionally included in the MCS for some datasets and/or loss functions. In particular, the MCS always contains the RT-GARCH model, and only for some datasets, the APARCH model with Student-t innovations. Moreover, the baseline RT-GARCH model always outperforms the standard GARCH(1,1) model for all horizons across all datasets. Hansen's

(2005) Test for Superior Predictive Ability (SPA) confirms the above results by showing that RT-GARCH model is not outperformed by any of the competing models. I also perform an evaluation of the forecasting performance of all models on two different subsamples: pre- and post-crisis periods. I show that during the crisis period, the RT-GARCH with leverage and RT-GARCH with leverage and feedback models outperform all other models for all stocks and all horizons. This result emphasises that during times of turmoil, accounting for leverage and especially allowing for a time-varying kurtosis is crucial for getting precise forecasts. Further, using VaR as an alternative risk measurement loss function, I show that my model has the correct conditional and unconditional coverage when compared to the other models, and especially when compared to the standard GARCH(1,1) model. Secondly, being a generalisation of the standard GARCH(1,1) model, RT-GARCH provides a better fit to the data when compared to the standard GARCH(1,1) model along several important dimensions. In particular, this is most evident in the tails of the standardised residual density implied by the estimated model. Lastly, I show how the RT-GARCH model can be used for specification testing of the standard GARCH models. This specification test can be interpreted as a test for constant conditional kurtosis against a time-varying one. Applied to IBM, GE and S&P500 data, I find that all of them have a time-varying conditional kurtosis.

The remainder of the chapter is structured as follows. In section 1.2 I introduce the RT-GARCH model and provide an interpretation of the model as well as its relation to GARCH and SV models. In section 1.3 I present the main results, including the conditional density function, and the strict and weak stationarity conditions. In section 1.4 I address the issue of leverage in the RT-GARCH model. Section 1.5 discusses some results of the estimation theory and the specification test. Section 1.6 describes how to use the RT-GARCH model to get l-step ahead volatility forecasts. In section 1.7 I provide an application to daily IBM, GE and S&P500 data. Section 1.8 concludes. All proofs are presented in the Appendix A.

1.2 RT-GARCH

1.2.1 Interpretation and relation to GARCH models

In this section I formally introduce the RT-GARCH model. In order to analyse the role of current information for volatility modelling, one first needs to define what is to be taken as "current information". Politis (2007) assumes that current information is represented by the current squared return. However, this poses a problem if one is to forecast the future conditional variance at time $t+1$ as the future return, r_{t+1}, will be required but is unobserved. One way to bypass this problem is to consider some function of the current return that won't require the knowledge of unobserved future returns when forecasting. One possible candidate for doing so is the current return scaled by its volatility. In GARCH-type models this translates directly into the error term, ϵ_t, which generates the return process. This solves the forecasting infeasibility issue as only the second conditional moment of the error term will be required for forecasting, which is known for all t, provided the standard moment conditions on the error term. More precisely, consider the following joint process $(r_t; \lambda_t^2)$:

$$r_t = \lambda_t \epsilon_t \tag{1.2}$$

$$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi \underbrace{\frac{r_t^2}{\lambda_t^2}}_{=\epsilon_t^2}, \qquad (\alpha, \beta, \gamma, \varphi) \geq 0, \tag{1.3}$$

where r_t is the return series, ϵ_t are i.i.d. random variables with a density function $f_\epsilon(\cdot)$ such that $E[\epsilon_t] = 0$, $E\left[\epsilon_t^2\right] = 1$. The true parameters are denoted by α_0, β_0, γ_0 and φ_0. This model nests the standard GARCH (1,1) model which can be obtained by setting $\varphi = 0$. I label the new volatility process λ_t^2 instead of σ_t^2, as eq.(1.3) does not correspond to the unconditional variance of returns r_t in this system of equations, i.e. $E[r_t^2] \neq E[\lambda_t^2]$ as λ_t is *not* independent of ϵ_t any longer. This makes the new model less tractable than the standard GARCH model. However, it can be shown that λ_t^2 is related to the conditional variance of returns as follows:

$$E\left[r_t^2|\mathcal{F}_{t-1}\right] = E\left[\lambda_t^2|\mathcal{F}_{t-1}\right] + \varphi\left(E\left[\epsilon_t^4\right] - 1\right). \tag{1.4}$$

On the other hand, the new specification can be seen as more flexible in certain dimension: it allows the shape of the conditional distribution of returns to vary over time - the feature that is absent from the standard GARCH models. Note also that the choice of a particular function of ϵ_t, i.e. ϵ_t^2, is only one of many possible ones subject to the necessary condition of $\lambda_t^2 > 0$. In particular, the functions $|\epsilon_t|$, ϵ_t^4 are possible. A reason for the decision to choose a squared error term will become apparent later when I discuss the interpretation and the implications for the conditional distribution of returns.

Although not directly related to the MIDAS approach of Ghysels et al. (2005, 2006), as I use only one frequency, the new model shares a similar intuition in the sense of assigning different and, in my case, *time-varying* weights to returns on different days. In particular, it can be shown that eq.(1.3) can be approximated by the following expression:

$$\lambda_t^2 = \frac{\varphi r_t^2}{b_{t-1}} + \sum_{j=1}^{\infty}\left(\frac{\beta^j\varphi}{b_{t-1-j}} + \gamma\beta^{j-1}\right) r_{t-j}^2 + \sum_{j=0}^{\infty} o\left(\frac{r_{t-j}^2}{b_{t-1-j}}\right), \tag{1.5}$$

where $b_{t-1} = \alpha + \beta\lambda_{t-1} + \gamma r_{t-1}^2$. The proof can be found in the Appendix A. Compared to the standard GARCH models, the weights are time-varying and depend on past volatility, which can be approximately taken to be b_{t-1}. The intuition of this weighting scheme is as follows. For the current return r_t^2, the weight is inversely proportional to b_{t-1}, i.e. the weight is bigger for a smaller past return and is smaller if the past return is large. For any r_{t-j}^2, $j \geq 1$, the weight consists of two parts: the usual "GARCH weight", given by $\gamma\beta^{j-1}$, and an additional time-varying weight $(\beta^j\varphi)/b_{t-1-j}$ which assigns an extra weight if a particular realisation of r_{t-j} is in the tails of the distribution.

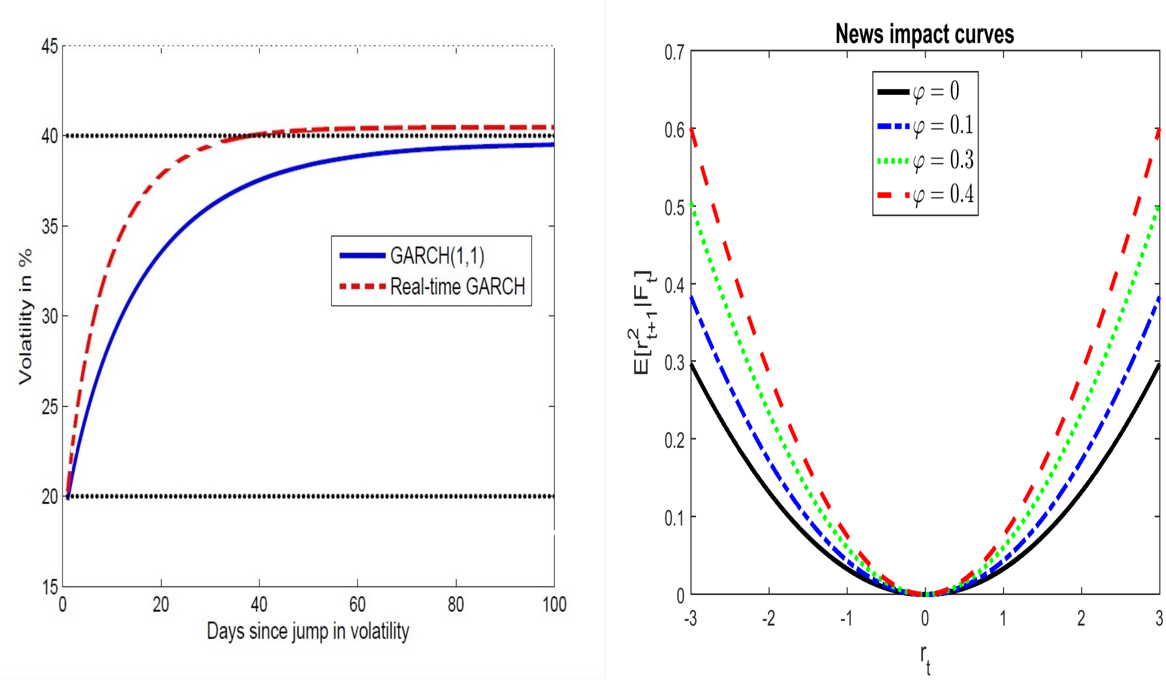
In order to understand the impact of enlarging the information content of the volatility process, consider the following thought experiment which I borrow from Hansen et al. (2012). Suppose that the true conditional variance σ_t^2 is such that the volatility is $\sigma_t = 20\%$ for $t \leq T$, but then suddenly jumps to the new level of $\sigma_t = 40\%$ for $t > T$. We then would like to investigate how a model-based conditional variance from GARCH and RT-GARCH models approximates σ_t^2, especially after a sudden jump. To answer this question we will calculate the expected conditional variance that is implied by each of the two models. For the standard GARCH(1,1) model for any $k \geq 0$ the expected conditional variance takes the form:

$$
\begin{aligned}
E\left[var(r_{T+k}|\mathcal{F}_{T+k-1})\right] &= E\left(\sigma_{T+k}^{2,GARCH}\right) = \alpha + \gamma E\left(r_{T+k-1}^2\right) + \beta\left[\alpha + \beta E\left(\sigma_{T+k-1}^{2,GARCH}\right) + \gamma E\left(r_{T+k-1}^2\right)\right] = \cdots = \\
&= \frac{\alpha}{1-\beta} + \gamma\sum_{j=0}^{\infty}\beta^j E\left(r_{T+k-1-j}^2\right) = \frac{\alpha}{1-\beta} + \gamma\sum_{j=0}^{k-1}\beta^j E\left(r_{T+k-1-j}^2\right) + \gamma\sum_{j=k}^{\infty}\beta^j E\left(r_{T+k-1-j}^2\right) = \\
&= \frac{\alpha}{1-\beta} + \gamma\frac{1-\beta^k}{1-\beta}(0.4)^2 + \gamma\frac{\beta^k}{1-\beta}(0.2)^2.
\end{aligned}
$$

Using similar derivation steps for RT-GARCH model with the important exception that $var(r_t|\mathcal{F}_{t-1}) = E\left[\lambda_t^2|\mathcal{F}_{t-1}\right] + \varphi\eta$, $\eta = E\left[\epsilon_t^4\right] - 1$, it similarly holds:

$$
E\left[var(r_{T+k}|\mathcal{F}_{T+k-1})\right] = \frac{\alpha + \varphi(3-2\beta)}{1-\beta} + \gamma\frac{1-\beta^k}{1-\beta}(0.4)^2 + \gamma\frac{\beta^k}{1-\beta}(0.2)^2,
$$

where I took $\epsilon_t \sim \mathcal{N}(0,1)$. In this thought experiment I ask the following question: how many days following the jump will it take for the volatility process to adjust to its new level? The answer is presented in Figure 1.1(a).



(a) (b)

Figure 1.1: (a) Time scale of the volatility adjustment. For both graphs the parameter vector $[\ \alpha,\ \ \beta,\ \ \gamma,\ \ \varphi]$ is set to $[\ 0,\ \ 0.92,\ \ 0.063,\ \ 0.035\ ]$ for the RT-GARCH, while $[\alpha, \beta, \gamma] = [0, 0.95, 0.045]$ for the standard GARCH(1,1) model. (b) News impact curves for different values of φ. For both graphs parameter values are estimated on the daily GE stock returns.

For the standard GARCH(1,1) model it takes approximately 100 days (more than 3 months) to approach the level of 39%. For the RT-GARCH it takes a little less than 40 days to adjust to the new level of volatility. Therefore RT-GARCH is at least two times faster in its speed of adjustment to the new level of volatility after a sudden jump when compared to the standard GARCH(1,1) model.

Another measure of how new information affects the volatility of returns is given by the "news impact curve", as defined by Engle and Ng (1993). For the RT-GARCH model the news impact curve is given by the following equation:

$$E\left[r_{t+1}^2|\mathcal{F}_t\right] = \alpha + \varphi\kappa + \beta\left(\frac{\bar{b} + \sqrt{\bar{b}^2 + 4\varphi r_t^2}}{2}\right) + \gamma r_t^2, \tag{1.6}$$

with $\kappa = E\left[\epsilon_t^4\right]$ and $\bar{b} = (\alpha + \beta\varphi + \kappa\gamma\varphi)/(1 - (\beta + \gamma))$ being the unconditional level of $b_{t-1} = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2$. Note that this news impact curve is no longer simply a quadratic function of r as in the case of standard GARCH(1,1) model. However, for reasonable values of the parameter values the last term in eq.(1.6) dominates. In Figure 1.1(b), I compare news impact curves of RT-GARCH for different values of φ with the news impact curve of the standard GARCH(1,1), which corresponds to the case of $\varphi = 0$. For a fixed value of φ the volatility in RT-GARCH model responds much more to extreme news when compared to the standard GARCH(1,1) model. For larger values of φ this response becomes even larger, see eq.(1.5) for the weighting interpretation. In the baseline model, however, good and bad news have the same weighting. I address the leverage and feedback issue and how it can be incorporated in the baseline model in Section 1.4.

1.2.2 Relation to SV models

To see how the new model relates to SV models I write the simplest possible versions of RT-GARCH and SV models, which is enough to demonstrate the point. Consider the following:

$$\left.\begin{aligned} r_t &= \lambda_t \epsilon_t \\ \lambda_t^2 &= \alpha + \varphi \epsilon_t^2, \\ \epsilon_t &\sim iid(0,1) \end{aligned}\right\} \quad RT-GARCH \qquad \left.\begin{aligned} r_t &= \sigma_t \eta_t, \\ \sigma_t^2 &= w + z_{t+1}^2, \\ z_t \sim iid(0,\sigma_z^2), \eta_t &\sim iid(0,\sigma_\eta^2) \\ \text{and } corr(z_{t+1},\eta_t) &= \rho \ \forall \ t \end{aligned}\right\} \quad \text{SV}$$

After simplifying both models as above, the difference becomes immediately clear. SV models assume that the process for returns, r_t, is driven by two random shocks, z_t and η_t. A non-zero contemporaneous dependence between shocks is allowed, which is thought to pick up the leverage effect, see Yu (2005) for the definition of leverage effect in SV models. Note that the inter-temporal dependence between shocks can be also allowed, see also Yu (2005) for a discussion, however this can lead to returns that are not martingale difference sequences and therefore not consistent with the efficient market hypothesis. The RT-GARCH model assumes that ϵ_t, a single random shock, is common to both r_t and its volatility process λ_t. My model is therefore a special case with $\rho = 1$ as the correlation of the shocks in the SV framework. As discussed above, this common shock only contributes to the volatility whenever it is large in absolute value. One therefore can think about it as really "bad" (in terms of both magnitude and sign) news that will be immediately incorporated in the volatility process. As mentioned before however, the RT-GARCH is neither a GARCH nor a SV model, but something in between. To formally define where in-between the new model lies, one would need to derive the continuous-time limit, which is left for future research.

1.3 Main Results

In this section I derive some statistical properties of the new model. I start with the unconditional moments of r_t^2 and λ_t^2. From eq.(1.2)-(1.3), the unconditional expectations of r_t^2 and λ_t^2 are given by:

$$E[r_t^2] = \alpha + \beta E[\lambda_{t-1}^2] + \gamma E[r_{t-1}^2] + \varphi E[\epsilon_t^4]$$

and

$$E[\lambda_t^2] = E[\alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi\epsilon_t^2] = \alpha + \beta E[\lambda_{t-1}^2] + \gamma E[r_{t-1}^2] + \varphi. \tag{1.7}$$

This now provides a link between the first moments of λ_t^2 and r_t^2:

$$E[r_t^2] = E[\lambda_t^2] + \varphi(E[\epsilon_t^4] - 1). \tag{1.8}$$

When, for instance, ϵ_t are i.i.d. $\mathcal{N}(0,1)$ random variables, the above relationship simply becomes $E[r_t^2] = E[\lambda_t^2] + 2\varphi$. I next derive the conditional density of returns together with the general formula for the jth conditional moment, followed by a discussion of the unconditional moments of r_t and λ_t^2. All proofs for this section's results can be found in the Appendix A.

Theorem 1. *Let ϵ_t be i.i.d. symmetric around zero random variables with density f_ϵ such that $E[\epsilon_t] = 0$ and $var(\epsilon_t) = 1$; and let (r_t, λ_t^2) evolve according to eq.(1.2)-(1.3). Denote by $\mathcal{F}_{t-1} := \sigma(r_t, s \leq t-1)$ the σ-algebra induced by the history of returns up to time $t-1$. Denote the parameter vector by $\theta = (\alpha, \beta, \gamma, \varphi)'$ and the true parameter vector by $\theta_0 = (\alpha_0, \beta_0, \gamma_0, \varphi_0)'$. Then the conditional probability density function of the return series, $f_r(r)$, is given by*

$$f_r(r|\mathcal{F}_{t-1}) = \frac{r}{d(r, b_{t-1}, \theta)\sqrt{b_{t-1}^2 + 4r^2\varphi}} f_\epsilon(d(r, b_{t-1}, \theta)), \tag{1.9}$$

where $f_\epsilon(\cdot)$ is the probability density function of ϵ_t, while $d(r, b_{t-1}, \theta)$ and b_{t-1} are given by the following equations

$$d(r, b_{t-1}, \theta) = \begin{cases} sign(r)\sqrt{\frac{\sqrt{b_{t-1}^2 + 4r^2\varphi} - b_{t-1}}{2\varphi}}, & for \quad \varphi \neq 0 \\ r/\sqrt{b_{t-1}}, \quad for \quad \varphi = 0 \end{cases} \tag{1.10}$$

with $b_{t-1} = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2$. Note that $\epsilon_t = d(r_t; b_{t-1}, \theta_0)$. Moreover,

$$\lim_{r\to 0} \frac{r}{d(r, b_{t-1}, \theta)} = \sqrt{b_{t-1}} \quad \text{and} \quad \lim_{r\to\ 0} f_r(r|\mathcal{F}_{t-1}) = \frac{1}{\sqrt{b_{t-1}}} f_\epsilon(0).$$

The conditional cumulative distribution function of returns is given by:

$$F(r|\mathcal{F}_{t-1}) = F_\epsilon\left(d(r, b_{t-1}, \theta)\right),$$

where $F_\epsilon(\cdot)$ is the cdf of ϵ_t. At θ_0 all odd conditional moments of r_t are zero and even conditional moments of r_t are given by the following formula:

$$E\left[r_t^j|\mathcal{F}_{t-1}\right] = b_{t-1}^{j/2}\sum_{i=-1}^{j/2-1}\frac{1}{(i+1)!}\left(\frac{\varphi}{b_{t-1}}\right)^{i+1}\left(\prod_{s=0}^{i}\left(\frac{j}{2}-s\right)\right)E\left[\epsilon_t^{j+2(i+1)}\right], \quad for \quad j \quad even. \tag{1.11}$$

In particular, from the formulae above it holds that

$$E\left[r_t^2|\mathcal{F}_{t-1}\right] = b_{t-1} + \varphi E\left[\epsilon_t^4\right],$$

and

$$E\left[r_t^4|\mathcal{F}_{t-1}\right] = b_{t-1}^2 E\left[\epsilon_t^4\right] + 2b_{t-1}\varphi E\left[\epsilon_t^6\right] + \varphi^2 E\left[\epsilon_t^8\right].$$

Remark 1. From eq. (1.11) it can be noticed that due to the symmetry assumption on the error term, all odd conditional moments of r_t will be zero, which then ensures that the returns are martingale difference sequence. Although definitely a stronger requirement than just assuming $E[\epsilon_t] = 0$, I believe it is still a realistic assumption as it will hold for a variety of distributions for ϵ_t. For instance, this requirement does not rule out the densities that are multimodal as long as they are still symmetric. In particular, it will hold for the commonly used Gaussian or Student-t distributions for ϵ_t.

Remark 2. Note that the distribution, related the conditional density in eq.(1.9) now has a time-varying kurtosis. Therefore, the parameter φ can be thought of as an extra shape parameter, representing the thickness of the tails. As a special case, this distribution nests the standard Normal distribution with a constant kurtosis of 3.

Remark 3. Conditional on $\mathcal{F}_{t-1}$, r_t is an odd function of ϵ_t, since ϵ_t is an odd function and λ_t is an even function of ϵ_t. Provided the symmetry condition on ϵ_t, it then automatically follows that the conditional, and hence unconditional distribution of r_t, is symmetric.

Remark 4. The conditional density of RT-GARCH model in eq.(1.9) nests the conditional density of the standard GARCH(1,1) model as its limiting case at $r = 0$. The intuition is as follows: standard GARCH(1,1) model is a special case of the RT-GARCH model whenever $\varphi = 0$, then $d(r)$ simplifies to $r/\sqrt{b_{t-1}}$ and eq. (1.9) boils down to the standard GARCH(1,1) density, or $\epsilon_t = 0$, which is equivalent to the condition of $r_t = 0$. In this case the limit of eq.(1.9) as $r \to 0$ is again the standard GARCH(1,1) density. Similarly, the conditional moments in eq.(1.11) nest the GARCH model conditional moments as its special case.

Remark 5. It is also interesting to note another important difference with the GARCH(1,1) model for conditional moments of order $j > 2$. Recall that for the standard GARCH(1,1) model the standardized conditional kurtosis of returns is just

$$E\left[r_t^4|\mathcal{F}_{t-1}\right]/\left(E\left[r_t^2|\mathcal{F}_{t-1}\right]\right)^2 = b_{t-1}^2 E\left[\epsilon_t^4\right]/b_{t-1}^2 = E\left[\epsilon_t^4\right],$$

meaning it is simply the standardised kurtosis of the error term, and therefore constant over time. For the RT-GARCH model, on the other hand, it holds that

$$E\left[r_t^4|\mathcal{F}_{t-1}\right]/\left(E\left[r_t^2|\mathcal{F}_{t-1}\right]\right)^2 = \frac{b_{t-1}^2 E\left[\epsilon_t^4\right] + 2\varphi b_{t-1} E\left[\epsilon_t^6\right] + \varphi^2 E\left[\epsilon_t^8\right]}{\left(b_{t-1} + \varphi E\left[\epsilon_t^4\right]\right)^2},$$

which makes it now time-varying. This explains why I opted to call φ an additional shape parameter, as it has a direct relationship to the standardised conditional kurtosis of the returns. In section 1.5 I discuss how this can be used for specification testing. Further note that the conditional distribution of the return series is no longer just the scaled version of the standard normal density. In particular, it now has an extra shape parameter φ, which, as I will describe below, will determine the degree of peakedness and/ or thickness of the tails of the distribution. In particular, the return process described by the RT-GARCH with normal innovations is now able to account for heavier tails compared to the standard normal distribution. To highlight this point further, Figures 1.2 and 1.3 display the probability density function of the RT-GARCH with $f_\epsilon(\cdot) \sim \mathcal{N}(0, 1)$ against the p.d.f. of the standard normal distribution and demonstrates that the density of the general model is also able to model heavier (than the standard normal) tails of the distribution without resorting to an arbitrary distribution of the error term.

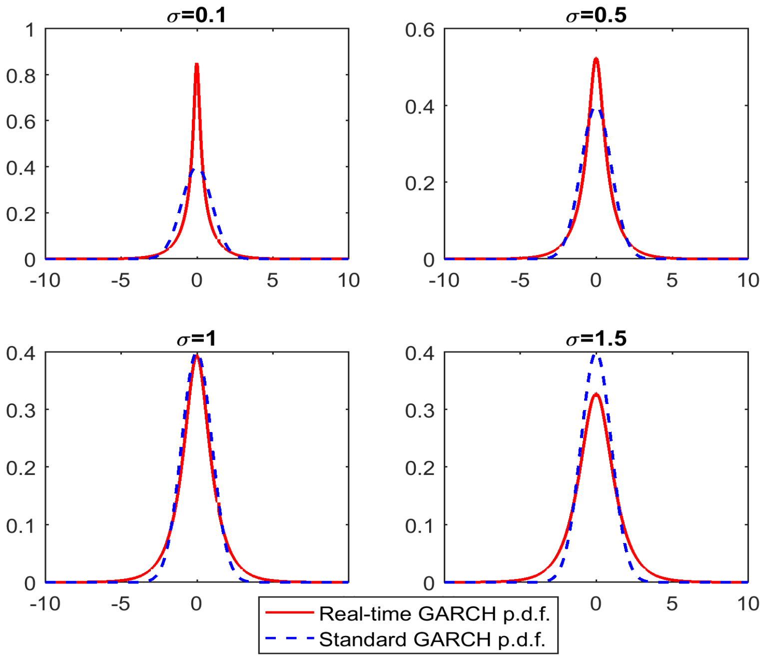


Figure 1.2: **Conditional probability density function for different values of unconditional volatility.** The parameter vector $\theta = [\ \alpha,\quad \beta,\quad \gamma,\quad \varphi]'$ is set to $[\ 0.003,\quad 0.9,\quad 0.04,\quad 0.02\ ]'$.

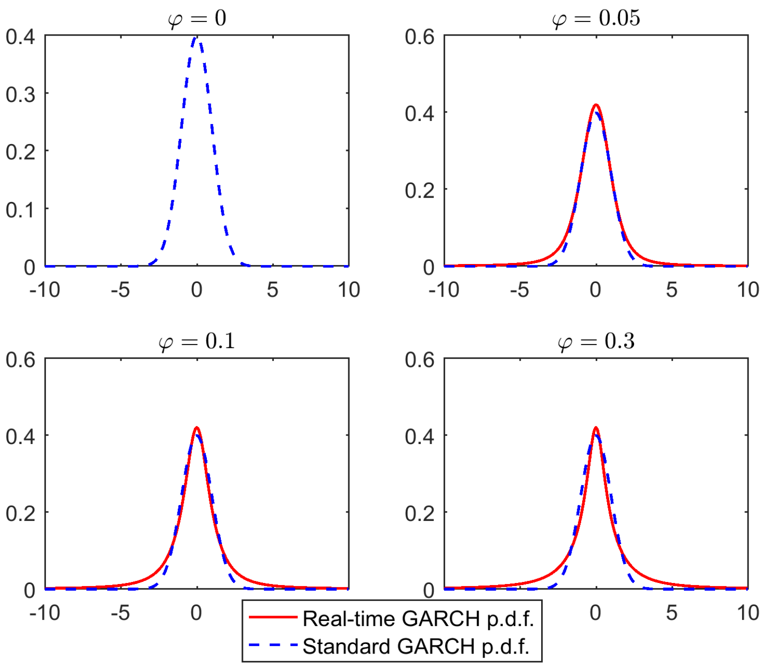


Figure 1.3: **Conditional probability density function for different values of φ.** The parameter vector $\theta = [\ \alpha,\quad \beta,\quad \gamma,\quad \varphi]'$ is set to $[\ 0.003,\quad 0.9,\quad 0.04,\quad 0.02\ ]'$.

The reason why the RT-GARCH model is able to reproduce heavy tails of the return's conditional distribution stems from the fact that the additional parameter φ controls the thickness of the tails of the corresponding distribution since the conditional kurtosis in the new model is time-varying. From the Figure 1.3 it is clear that the larger the value of φ is, the heavier are the tails of the distribution. Besides controlling for the thickness of the tails of the distribution, the parameter φ allows for the adjustment of the volatility estimate, either up or down depending on the "sign" (i.e. positive or negative) of the news, allowing the conditional variance process to be more dynamic.

I now turn to describing some further statistical properties of the RT-GARCH model. In particular, I derive conditions for the joint process (r_t, λ_t^2) to be strictly stationary. This result will be further used in developing the asymptotic theory for the quasi-maximum likelihood estimator (QMLE) of the parameter vector, which I present in Chapter 2.

Theorem 2. *Let ϵ_t be i.i.d. symmetric around zero random variables with density f_ϵ such that $E[\epsilon_t] = 0$ and $var(\epsilon_t) = 1$; and let (r_t, λ_t^2) evolve according to eq.(1.2)-(1.3). Let $\beta, \gamma > 0$, and $\varphi \neq 0$. If the following conditions are satisfied*

$$-\infty \leqq E \log \left|\beta + \gamma\epsilon_0^2\right| < 0 \qquad E\left(\log\left|\alpha + \varphi\epsilon_0^2\right|\right)^+ < \infty, \tag{1.12}$$

then the process (r_t, λ_t^2) is strictly stationary.

I next establish the weak stationarity conditions for r_t^2 and λ_t^2 processes. These results will be later used to derive the forecasting formulae for the conditional variance of returns. In addition, the unconditional level of volatility is needed if one chooses to use variance targeting for the estimation of the parameter vector.

Theorem 3. *Let ϵ_t be i.i.d. symmetric around zero random variables with density f_ϵ such that $E[\epsilon_t] = 0$, $var(\epsilon_t) = 1$ and $E[\epsilon_t^4] < \infty$; and let (r_t, λ_t^2) evolve according to eq.(1.2)-(1.3). Then if $\beta + \gamma < 1$ the process λ_t is weakly stationary and its second unconditional moment is given by*

$$E[\lambda_1^2] = \frac{\alpha + \varphi + \gamma\varphi\left(E[\epsilon_t^4] - 1\right)}{1 - (\beta + \gamma)}. \tag{1.13}$$

Given the relationship, described in eq.(1.8), between $E[r_t^2]$ and $E[\lambda_t^2]$, it is possible now to write down the conditions for weak stationarity of r_t^2.

Theorem 4. *Let ϵ_t be i.i.d. symmetric around zero random variables with density f_ϵ such that $E[\epsilon_t] = 0$, $var(\epsilon_t) = 1$, and $E[\epsilon_t^4] < \infty$; and let (r_t, λ_t^2) evolve according to eq.(1.2)-(1.3). Then if $\beta + \gamma$, r_t is weakly stationary and its second unconditional moment is given by*

$$E[r_1^2] = \frac{\alpha + \varphi E[\epsilon_t^4] + \varphi\beta\left(1 - E[\epsilon_t^4]\right)}{1 - (\beta + \gamma)}. \tag{1.14}$$

In addition it also holds that:

$$cov(r_t, r_s) = 0, \qquad t \neq s.$$

I now turn to the unconditional fourth moment of the return series r_t, $E[r_1^4]$, which is an important measure of the tail behaviour of the return distribution. Recall that the process $\{X_t\}$ is fourth-moment stationary if $E\left[X_{t_1+k}X_{t_2+k}X_{t_3+k}X_{t_4+k}\right] = E\left[X_{t_1}X_{t_2}X_{t_3}X_{t_4}\right]$ for all values of k and all values of t_1, t_2, t_3 and t_4. Detailed derivations are presented in the Appendix A. Here I present the resulting expression for $E[r_1^4]$.

Theorem 5. *If the process (r_t, λ_t^2) evolves according to eq. (1.2)-(1.3) and ϵ_t are symmetric around zero i.i.d. random variables with density f_ϵ such that $E\left[\epsilon_t\right] = 0$ and $var(\epsilon_t) = 1$, and $E[\epsilon_t^8] < \infty$ then r_t is fourth-moment stationary if*

$$\gamma^2 < \frac{1}{E[\epsilon_t^4]}, \tag{1.15}$$

with the unconditional fourth moment given by

$$E[r_1^4] = \frac{\xi_1 + E[\lambda_1^2]\xi_2 + \left(\beta^2 + 2\beta\gamma\right)\mu_4\left[E[\lambda_1^2]\right]^2}{1 - \gamma^2\mu_4}, \tag{1.16}$$

where $\mu_j := E[\epsilon_t^j]$, $\eta := E\left[\epsilon_t^4\right] - 1$, $\chi := \beta^2/(1-\beta^2)$ and constants ξ_1 and ξ_2 are given by $\xi_1 = \left(\alpha^2 + 2\eta\alpha\gamma\varphi + \varphi^2\chi\eta\right)\mu_4 + \varphi^2\mu_8 + 2\varphi\mu_6\left(\alpha + \eta\varphi\gamma\right) > 0$ and $\xi_2 = 2\mu_4(\alpha\beta + \alpha\gamma + \eta\beta\gamma\varphi) + 2\mu_6(\beta\varphi + \gamma\varphi) > 0$ and $E[\lambda_1^2]$ is given by eq. (1.13).

Remark 6. In the case of Gaussian error terms, condition (1.15) simply becomes $\gamma^2 < \frac{1}{3}$ which is exactly the same as in the standard GARCH(1,1) case.

1.4 Leverage and volatility feedback effects

The RT-GARCH model described by eq. (1.2)-(1.3) has no leverage effect, meaning that when errors are symmetric about zero, $E[r_t] = 0$ and $cov(r_t^2, r_j) = 0 \quad \forall j$. However, there is well documented empirical evidence, see e.g. Black (1976), Christie (1982), Engle and Ng (1993), that many financial time series exhibit the leverage effect, i.e. the contribution to the volatility of negative shocks to the stock prices is far greater than that of the positive shocks of the same magnitude. As a result of this empirical evidence, most discrete and continuous-time volatility models were extended to incorporate this feature. For discrete time models see Nelson (1991), Engle and Ng (1993), Glosten et al.(1993) among others. For continuous-time models, see Christie (1982), Yu (2005), Bandi and Renò (2012), Aït-Sahalia et al. (2013) and Wang and Mykland (2014). For a fully nonparametric way of estimating and testing the leverage hypothesis, see also recent work by Linton et al. (2016).

I proceed by incorporating the leverage effect in the fashion of Glosten et al. (1993), i.e. by acknowledging the different effect of positive and negative news on the conditional variance of returns. Note however that, unlike for the standard GARCH-type models, the most recent information in my case is represented by current shocks ϵ_t. I therefore refer to "leverage effect" by differentiating the effect of positive and negative values of ϵ_t on λ_t^2. Therefore the baseline model in Section 1.2 can be extended to account for leverage effect as follows:

$$r_t = \lambda_t \epsilon_t$$

and

$$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi_1 \epsilon_t^2 \mathbb{1}_{(\epsilon_t > 0)} + \varphi_2 \epsilon_t^2 \mathbb{1}_{(\epsilon_t \leq 0)}.$$

It is also interesting to differentiate between the effect of positive and negative values of past returns on the conditional volatility. In the standard GARCH-type models this is referred as "leverage effect" as this would be the most recent information effecting the conditional volatility. Given the differently defined leverage effect in my model I refer to the different effects of the past positive and negative returns on conditional variance as "feedback effect". More precisely, the RT-GARCH model with leverage and feedback effects is given by

$$r_t = \lambda_t \epsilon_t \tag{1.17}$$

and

$$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma_1 r_{t-1}^2 \mathbb{1}_{(r_t > 0)} + \gamma_2 r_{t-1}^2 \mathbb{1}_{(r_t \leq 0)} + \varphi_1 \epsilon_t^2 \mathbb{1}_{(\epsilon_t > 0)} + \varphi_2 \epsilon_t^2 \mathbb{1}_{(\epsilon_t \leq 0)}. \tag{1.18}$$

In Figure 1.4 I compare the news impact curves of the GJR-GARCH(1,1) of Glosten et al.(1993) with the RT-GARCH with leverage and RT-GARCH with leverage and feedback, all estimated on the daily IBM data. For both specifications of the RT-GARCH model, volatility tends to respond more to negative news than in GJR-GARCH model. Note, however, that the RT-GARCH model with leverage and feedback responds slower to negative news than the RT-GARCH just with the leverage effect.

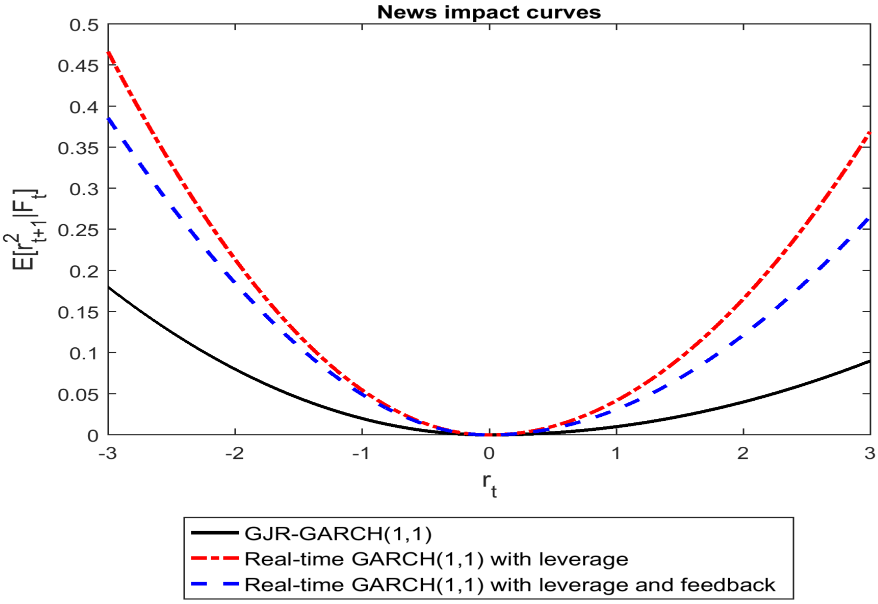


Figure 1.4: The figure displays the news impact curves for three models, estimated on the daily IBM data.

All theorems in section 1.3 hold for both extensions with slight modifications. For reasons of brevity I defer these to the online Supplementary Material for this chapter.

1.5 Outline of the Estimation Theory

In this section I discuss some results of the QMLE analysis. I denote the parameter vector by $\theta = (\alpha, \beta, \gamma, \varphi)'$ and the corresponding true parameter vector by $\theta_0 = (\alpha_0, \beta_0, \gamma_0, \varphi_0)'$. For the purpose of estimation I adopt a Gaussian specification, such that the log-likelihood function can be written as follows:

$$L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} l_t(\theta),$$

where

$$l_t(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}d_t^2(r_t, b_{t-1}, \theta) + \log\left(\frac{\sqrt{b_{t-1}(\theta) + \varphi d_t^2(r_t, b_{t-1}, \theta)}}{b_{t-1}(\theta) + 2\varphi d_t^2(r_t, b_{t-1}, \theta)}\right),$$

with $b_{t-1}(\theta) = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2$, and $d_t^2(r, b_{t-1}, \theta)$ is given in eq.(1.10). Note also that if φ is set to zero we are again back to the standard GARCH(1,1) log-likelihood function. The estimator of interest θ_0 is then defined as follows:

$$\widehat{\theta}_T = \arg\max_{\theta\in\Theta} \widetilde{L}_T(\theta), \tag{1.19}$$

where Θ denotes the parameter space. If the error term $\epsilon_t := d(r_{t,t-1}, \theta_0)$ is Gaussian, then $\widehat{\theta}$ is MLE, otherwise it is a QMLE. Given that the RT-GARCH(1,1) model nests the standard

GARCH(1,1) model, it can be expected that the asymptotic theory for QMLE will be a generalisation of some sort for the standard GARCH(1,1) model. In fact, this turns out to be true, however the analysis is non-trivial and requires lengthy derivations. The entire analysis presents an interest of its own, and therefore it is studied in detail in Chapter 2. Here I provide only a brief discussion of the results. In particular, in Chapter 2 I show that the joint process (r_t^2, λ_t^2) is ergodic, and establish strong consistency of $\widehat{\theta}_T$ by adopting the theory by Francq and Zakoïan (2004). In addition, I also show that the score function is still a martingale difference sequence, therefore the martingale CLT, see e.g. Hall and Heyde (1980), can be applied to show:

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, V_\theta\right),$$

where $V_\theta \equiv A^{-1}BA^{-1}$ and

$$A = -\frac{1}{T}E_{\theta_0}\left[\frac{\partial^2 \log L_T(\theta)}{\partial\theta\partial\theta'}\right] \qquad \text{and} \qquad B = \frac{1}{T}E_{\theta_0}\left[\frac{\partial \log L_T(\theta)}{\partial\theta}\frac{\partial \log L_T(\theta)}{\partial\theta'}\right].$$

The exact expressions for V_θ are presented in Chapter 2, where the theory is studied at length. Finally, provided that $\widehat{\theta} \xrightarrow{p} \theta_0$ and $\widehat{V}_\theta \xrightarrow{p} V_\theta$, the feasible version is given by:

$$\widehat{V}_\theta^{-1/2}\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, I\right).$$

I finish this section by suggesting that the new model can be used for specification testing of the standard GARCH models. In particular, one can consider testing of the following null hypothesis:

$$\mathbb{H}_0: \qquad \varphi = 0$$

versus an alternative hypothesis $\mathbb{H}_A$ that $\mathbb{H}_0$ is false. This test can be interpreted as the test for constant standardised conditional kurtosis of the returns against an alternative of a time-varying conditional kurtosis. Since this test is for nested models, it is straightforward to use already computed likelihood quantities to calculate the Likelihood Ratio test $LR = -2\ln(\widetilde{L}_T(\theta^\star)/L_T(\theta)) \xrightarrow{d} \chi_1^2$, where $\theta = (\alpha, \beta, \gamma, \varphi)'$, $\theta^\star := \{\theta \setminus \varphi\}$. Although theoretically φ can take negative values, see Theorems 3 and 4 for restrictions, in practical applications the easiest way to ensure that λ_t^2 is always positive is to restrict all parameters to be positive, i.e. $\varphi \geq 0$. In this case the test is on the boundary of the parameter space for φ, and the Likelihood ratio test has a nonstandard distribution, see Francq and Zakoïan (2009) for details.

1.6 Volatility Forecasts with RT-GARCH

I now focus on volatility forecasting using the RT-GARCH model. The forecasting exercise is very similar to obtaining volatility forecasts with the standard GARCH(1,1) model except for some slight differences. Recall that for the forecasting exercise the following two equations are needed:

$$E[\lambda_t^2|\mathcal{F}_{t-1}] = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi \tag{1.20}$$

and

$$E[r_t^2|\mathcal{F}_{t-1}] = E[\lambda_t^2|\mathcal{F}_{t-1}] + \varphi(E[\epsilon_t^4] - 1), \tag{1.21}$$

where eq. (1.20) is the expectation of the conditional variance of the process and eq. (1.21) is obtained by recursively substituting eq. (1.3) into the squared eq. (1.2) and taking expectations. Then k-step ahead volatility forecast formulae are given in Theorem 6 below.

textbfTheorem 6. *Let the process* (r_t, λ_t^2) *evolve according to eq. (1.17)-(1.18) and* ϵ_t *are symmetric around zero i.i.d. random variables with density* f_ϵ *such that* $E[\epsilon_t] = 0$ *and* $var(\epsilon_t) = 1$. *Moreover, let the estimator* $\widehat{\theta} = \left(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}_1, \widehat{\gamma}_2, \widehat{\varphi}_1, \widehat{\varphi}_2\right)'$ *be the QMLE estimator for* $\theta = (\alpha, \beta, \gamma_1, \gamma_2, \varphi_1, \varphi_2)'$ *in the extended model defined in eq. (1.18). Then the k-step ahead,* $k \geq 1$, *volatility forecast is given by the following formula:*

$$E[r_{t+k}^2|\mathcal{F}_t] = E[\widehat{\lambda}_{t+k}^2|\mathcal{F}_t] + (\widehat{\varphi}_1 + \widehat{\varphi}_2)(E[\epsilon_{t+k}^4|\mathcal{F}_t] - 1) = \widehat{\Lambda} + (\widehat{\beta} + \widehat{\gamma}_1 + \widehat{\gamma}_2)^k \left(E[\widehat{\lambda}_t^2|\mathcal{F}_t] - \widehat{\Lambda} \right) +$$
$$+ (\widehat{\varphi}_1 + \widehat{\varphi}_2)(E[\epsilon_{t+k}^4|\mathcal{F}_t] - 1),$$

where $\widehat{\Lambda}$ *is given by*

$$\widehat{\Lambda} = \frac{\widehat{\alpha} + (\widehat{\varphi}_1 + \widehat{\varphi}_2)\left[\eta(\widehat{\gamma}_1 + \widehat{\gamma}_2) + 1\right]}{1 - \left(\widehat{\beta} + \widehat{\gamma}_1 + \widehat{\gamma}_2\right)},$$

with $\eta \equiv E\left[\epsilon_t^4\right] - 1$ *and* $E\left[\widehat{\lambda_t^2}|\mathcal{F}_t\right]$ *is calculated by using* $\widehat{\theta}$ *as an estimator of* θ.

Note that Theorem 6 provides the most general formulae for the RT-GARCH with leverage and feedback. Forecasting formulae for the RT-GARCH model with leverage only may be obtained by setting $\gamma_1 = \gamma_2 = \gamma$, while for the baseline RT-GARCH model by setting $\gamma_1 = \gamma_2 = \gamma$ and $\varphi_1 = \varphi_2 = \varphi$.

1.7 Application

1.7.1 Data and Methodology

To estimate and evaluate competing models I use 3 datasets of open-to-close returns, namely IBM, General Electric (GE) and the S&P 500 index. The original data was bought from Kibot and constitutes DJ30 1-minute high-frequency return data, of which I picked IBM and GE for my empirical application. The data was then aggregated to 5-minutes for calculating the 5-minute realised volatility to serve as a proxy of the end-of-the day volatility. I also use the SP500 index historical returns which is freely available from the realised library at Oxford-Man Institute of Quantitative Finance. This data includes the open-to-close daily returns and the 5-minute realised volatility. For IBM and GE the data spans from the 2nd January 1998 until the 1st December 2016, while for the S&P500 the time span is from the 28th January 2003 until the 1st December 2016. In order to avoid estimation bias, I split the sample into two parts, the first part will serve for the model's estimation, and the remaining part will be used for an evaluation of the out-of-sample performance using the recursive scheme. As with any out-of-sample forecasting exercise there is no direct guidance of the optimal splitting point. For presenting the main results, I reserve 2/3 of the whole sample for the estimation and the rest of the sample for the forecast evaluation. For the IBM and GE stocks this results in 3000 and 1500 observations for estimation and evaluation respectively. For the S&P500 data I have 2000 and 1000 observations for estimation and evaluation respectively. However, in order to make sure that the results do not depend on the splitting point, in the online Supplementary Material to this chapter I present results for different splitting points. The issue of the splitting point and how it affects forecast evaluation tests will be further addressed in Chapter 3. In addition to the full sample results, due to the likelihood of structural breaks during the financial crisis period, I also present results for 2 subsamples: pre- and post-crisis periods. The pre-crisis period spans from 2nd of January (28th of January 2003 for the S&P 500 index) till the end of July 2008. The crisis and post-crisis period constitutes the rest of the available sample. For subsamples for GE and IBM stocks I take 1500 and 500 observations for estimation and evaluation respectively. For subsamples for S&P500 data I use 1000 and 500 observations for estimation and evaluation respectively.

For out-of-sample forecast performance I compare RT-GARCH models with the standard GARCH(1,1), GARCH(1,2) with Normal and Student-t innovations, APARCH model with Student-t innovations (as the most sophisticated GARCH-type model, see Hansen and Lunde (2005) for details), as well as Simple and Exponential NoVaS methodologies of Politis (2007). The specifications of all competing models are presented in Table 1.1. I exclude SV models from this comparison as SV models are outside of the Maximum Likelihood framework. Moreover, since the purpose of this chapter is not to propose *the* best volatility model but rather investigate whether incorporating available current information in GARCH-type models will improve on existing GARCH models in terms of out-of-sample volatility and VaR forecasts, inclusion of SV models is not necessary to answer this question.

The "true volatility" would be needed in order to directly evaluate the forecasting performance of competing models. Without the true volatility process, the most common approach instead is to use realised volatility as a proxy for the conditional variance of returns. I calculate the 5-minute realised variance from the intraday high-frequency data for each stock, which I then take to be the proxy for the conditional variance of returns in out-of-sample forecast evaluations.

1.7.2 Results and Discussion

In this section I report the parameter estimates for the RT-GARCH, RT-GARCH with leverage effect (RT-GARCH-L) and RT-GARCH with leverage and feedback effect (RT-GARCH-LF). Results are presented in Table 1.2. For all RT-GARCH models and all datasets the parameter φ is positive and significantly different from zero. Note that for the model with leverage, the value of the parameter φ_2 is much larger than the value of the parameter φ_1, pointing at the fact that negative news contribute to volatility more than positive ones.

For out-of-sample evaluation I use the only two "robust" loss functions (see Patton, 2011) in the context of volatility forecasting. Note that a loss function is "robust" if for any two volatility forecasts, h_{1t}^2 and h_{2t}^2, their ranking according to expected loss is equivalent whether it is done using the true conditional variance, σ_t^2, or some proxy $\hat{\sigma}_t^2$, provided the latter is conditionally unbiased, i.e. $E[r_t^2|\mathcal{F}_{t-1}] = E\left[\hat{\sigma}_t^2|\mathcal{F}_{t-1}\right] = \sigma_t^2$.

Tables 1.3-1.14 present the results. For the presentation of results I adopt the original notation of Hansen et al. (2011), i.e. $\widehat{\mathcal{M}}^{\star}_{95\%}$ denotes the MCS $\widehat{\mathcal{M}}^{\star}$ that contains the best models with probability 0.95. For both statistical loss functions, MSE and QLIKE, Real-time-GARCH and RT-GARCH-L models are always in the MCS $\widehat{\mathcal{M}}^{\star}_{95\%}$ for all horizons, while standard GARCH models most of the time fall outside of the MCS. I present the results for full sample as well as the results for pre- and post-crisis (including crisis) subsamples.

I start with the full sample results. For the 1-step ahead out-of-sample volatility forecasts using the MSE loss function, the MCS for the IBM stock is quite wide and consists of all competing models except for the NoVaS methodologies, while for the QLIKE loss function the MCS consists solely of all the RT-GARCH models. For the GE stock for 1-step ahead forecasts MCS consists of all RT-GARCH models and the APARCH(2,2) model for both loss functions. Finally, for the S&P 500 stock the MCS based on the MSE loss function is quite small and consists only of RT-GARCH and RT-GARCH-L models, while the MCS based on the QLIKE loss function consists of RT-GARCH, RT-GARCH-L and APARCH(2,2) models.

For the 5-step ahead forecasts, the picture is very similar, except that for the MSE loss function, the MCS sometimes includes the GARCH models with Student-t innovations. For example, for the 5-step ahead forecasts using IBM data, for the MSE loss function the MCS consists only of RT-GARCH model, while for the QLIKE loss function both GARCH models with Student-t innovations are included as well. A similar picture can be seen for the GE stock for the MSE loss function, while for the QLIKE loss function the MCS consists again only of RT-GARCH, RT-GARCH-L and the APARCH(2,2) models. For the S&P 500 stock for the MSE loss function the MCS consists of all competing models but NoVaS methodologies, while for the QLIKE loss function the MCS consists only of the RT-GARCH and APARCH(2,2) models.

For longer horizons, i.e. 10- and 15-step ahead out-of-sample volatility forecasts, the picture is quite different. More precisely, the MCS consists only of the RT-GARCH and the APARCH(2,2) models, with the occasional inclusion of RT-GARCH-L model and sometimes GARCH models with Student-t innovations.

Note that for all horizons, the standard GARCH models with Gaussian innovations are excluded from the MCS for all stocks. It is also interesting to note that most of the time, MCS for all datasets always contain Student-t innovations (which allows for heavier tails) and RT-GARCH models. Note however that RT-GARCH models perform no worse (or most of the time even better) with just the normal innovations. As discussed in section 1.2 the possible reason for this is that the RT-GARCH models account for a time-varying conditional kurtosis, therefore allowing the volatility to adjust to a new level faster than the other standard GARCH models. It is also possible that the forecasting performance of RT-GARCH models can be further improved if one considers Student-t innovations for the error term. On the other hand, RT-GARCH model with leverage and feedback effects (RT-GARCH-LF) seems to perform worse than the simple RT-GARCH or RT-GARCH-L, as it can potentially overfit the data due to the model's higher complexity (i.e. higher number of parameters).

Given that all samples under consideration include the financial crisis, it is important to account for the structural break in the volatility of returns. If one is to account for the structural break, the parameters of each model have to be re-estimated during/after the break. I address this issue by estimating and evaluating the models on two subsamples: pre- and post-crisis period, where the latter includes the crisis period as well.

While the forecast evaluation results for the pre-crisis period are quite similar to the full sample results, the crisis period MCS is quite different for all stocks. For the crisis and post-crisis period the MCS for both loss functions mainly consists of RT-GARCH-L, RT-GARCH-LF and the APARCH(2,2) models. This result is general for all stocks and all horizons. The difference in results emphasises that during volatile periods it is crucial to account for both leverage and time-varying kurtosis.

There are several reasons why NoVaS methodologies are never in the MCS. First of all, Politis (2007) compares forecasts with the Mean Absolute Deviation (MAD) loss function, which is not a robust loss function in the context of volatility forecasting, see Patton (2011). The other reason may be that the comparison of NoVaS forecasts was done with the use of squared returns as a volatility proxy, which was shown to be a quite noisy proxy in the context of volatility forecasting, see e.g. Hansen and Lunde (2006).

In addition, I also evaluate all forecasts with the risk management loss function, i.e. I compute 1-step Value-at-risk (VaR) forecasts using all competing models. For evaluation of VaR forecasts I compute the Violation Ratio (VR), which is the ratio between the number of returns that exceeded the VaR forecast to the number of the expected exceedances, accounting for a significance level of α which I take to be 5%. If the model is accurate, the violation ratio is expected to be exactly 1. A model has good forecasts if the VR is between 0.8 and 1.2; and a model has quite imprecise forecasts if the VR<0.5 or VR>1.5. However, computing only the VR is not enough for evaluating VaR forecasts as it is the measure of the unconditional coverage. I therefore also compute the Likelihood Ratio (LR) for the conditional coverage from the failure process of the VaR forecasts, see Christoffersen (1998) for details. Table 1.15 presents the results. Out of all models with a correct conditional coverage, RT-GARCH (for all stocks) and RT-GARCH with leverage (for IBM stock) are the only models that have an acceptable VR. In addition, this ratio will be far better than for the standard GARCH(1,1) model with normal errors for all stocks under consideration. This result further emphasises the effect of having a time-varying kurtosis of returns, which allows for the possibility of adjusting it over time in response to the data, playing a potentially crucial role for forecasting of VaR.

After identifying which models are in the MCS, it is still interesting whether it is possible to pin down a single superior forecasting model (in the sense that it is not outperformed by any other competing model) among those in the MCS. One possibility is to conduct an out-of-sample test that has the ability to control either for possible over-fitting or over-parametrisation problems, which gives a more powerful framework to evaluate the performances of competing models. I choose to conduct Hansen's (2005) Test for Superior Predictive Ability (SPA). Results of the test for all models are produced for one-step ahead volatility forecasts and for reasons of brevity these are presented in Appendix B. The overall conclusion is that the winning model (among those in the MCS) is one of the RT-GARCH models for shorter horizons (i.e. 1- and 5-step ahead) and either APARCH model or RT-GARCH/RT-GARCH-L for longer horizons.

In addition, I perform the likelihood ratio test for $H_0: \ \varphi = 0$, adjusted for testing on the boundary, see Francq and Zakoïan (2009) for details. The values of the test statistic are 8.5, 4.66 and 9.72 for IBM, GE and S&P500 respectively, which are significant at a 5% significance level.

This suggests that all time series have a time-varying conditional kurtosis.

In order to show that the RT-GARCH model is a better fit to the data, especially in the tails, Figures 1.5-1.7 displays the QQ plots of the standardised errors from the estimated GARCH(1,1) and RT-GARCH models.

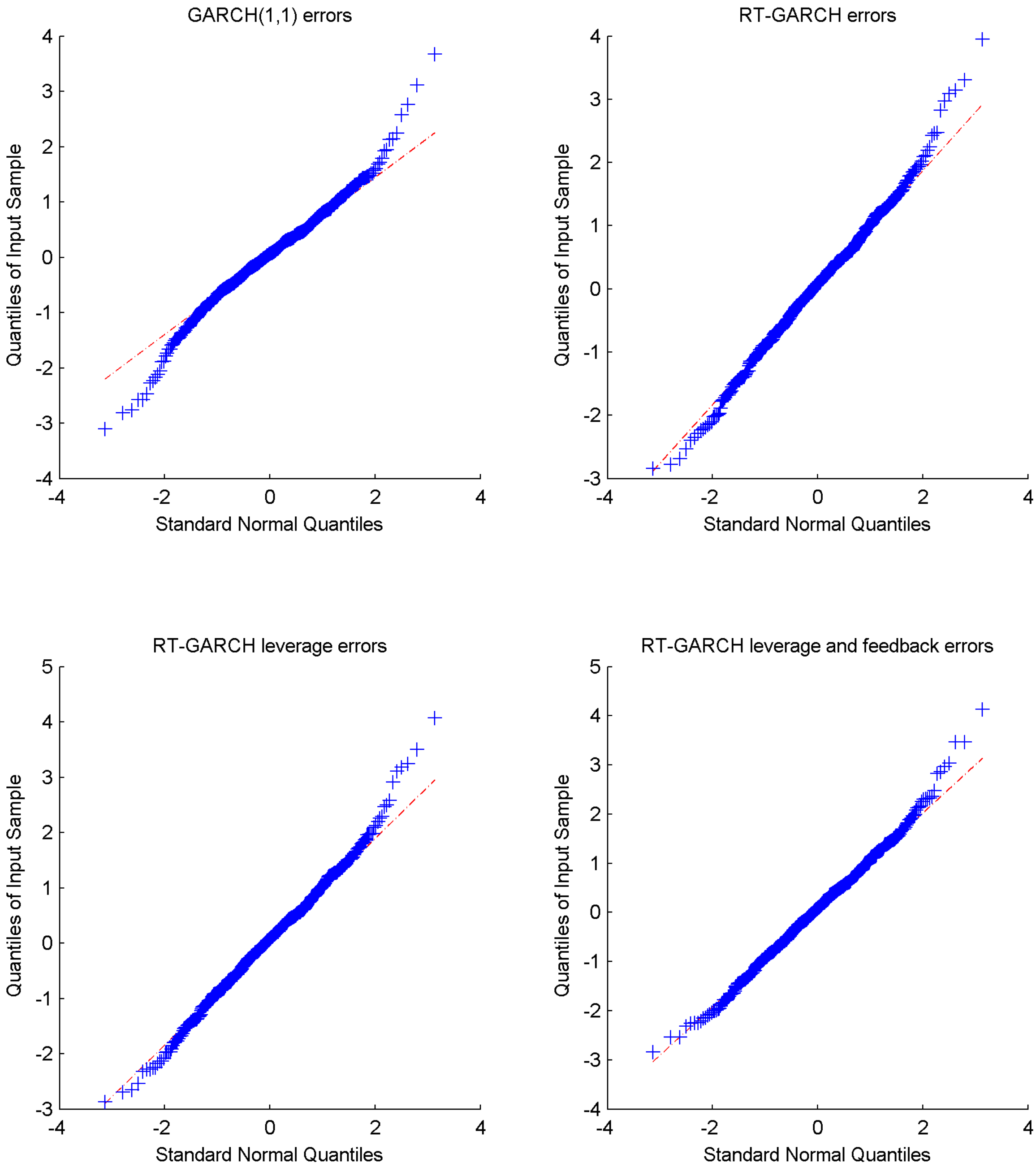


Figure 1.5: QQ-plots of the implied error distribution for IBM stock.

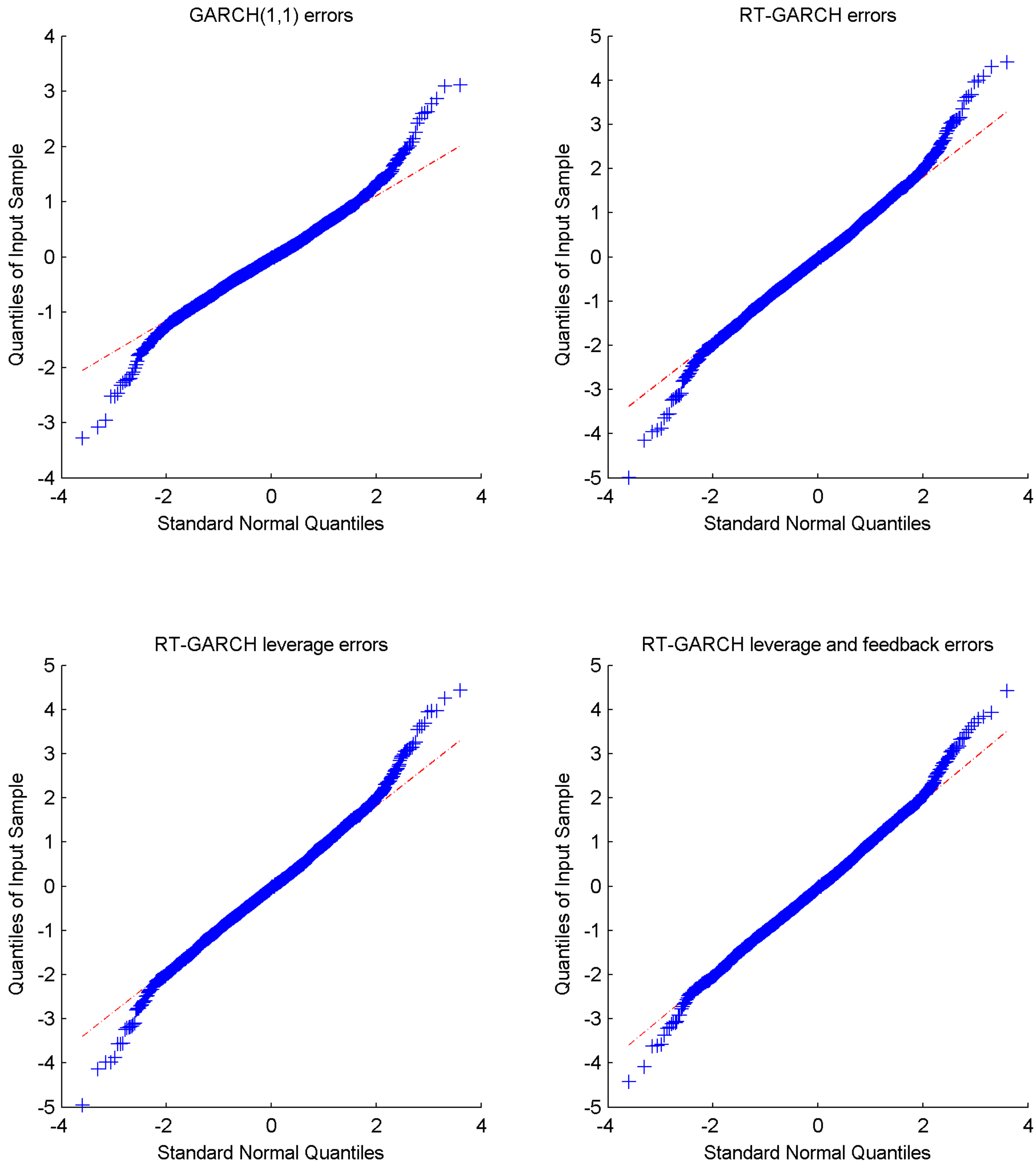


Figure 1.6: QQ-plots of the implied error distribution for GE stock.

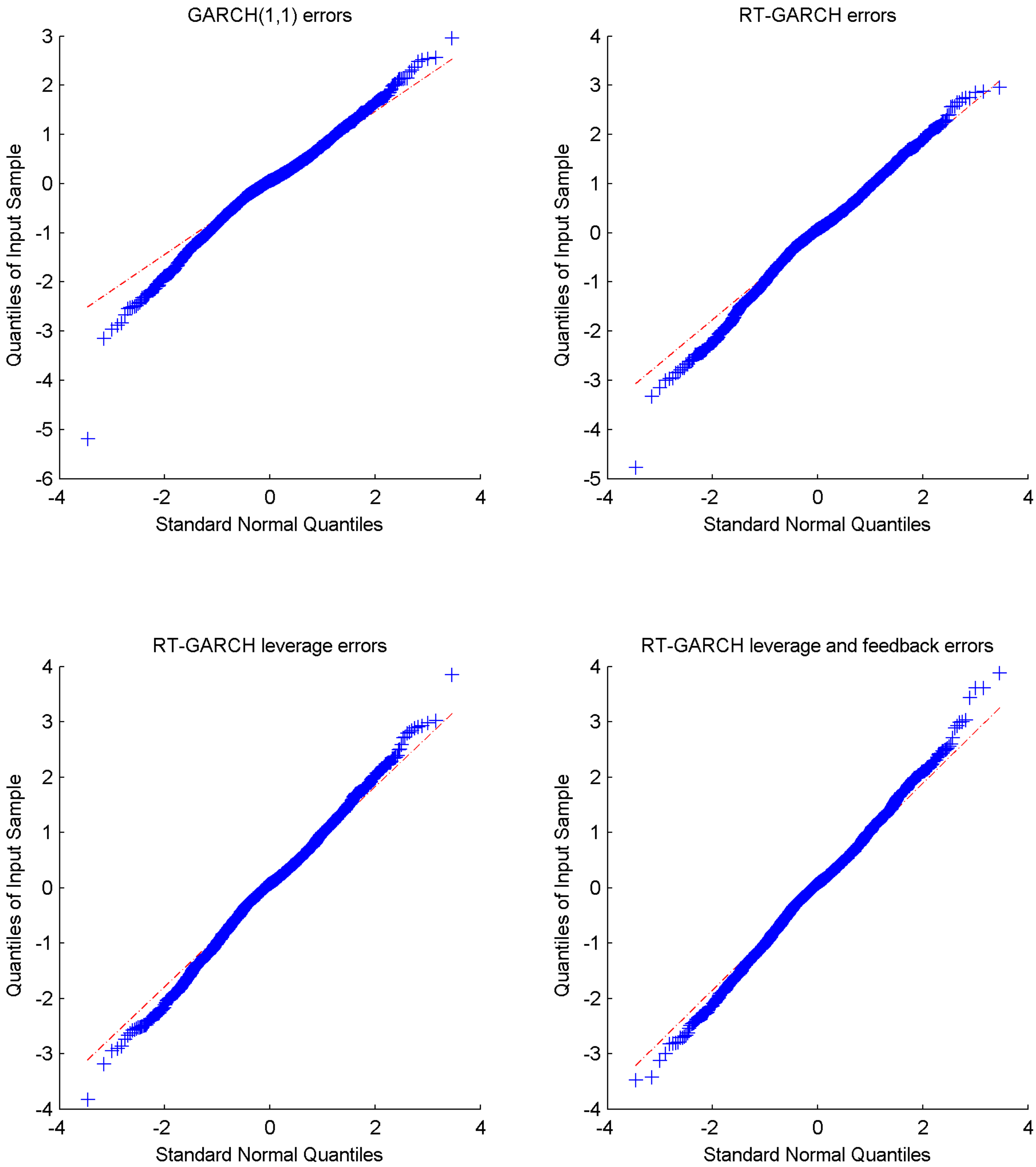


Figure 1.7: QQ-plots of the implied error distribution for S&P500 index.

1.8 Conclusion

Volatility of asset returns is difficult to forecast due to its latent nature. In an attempt to describe the volatility process the majority of the discrete-time volatility models incorporate only past information in modelling the volatility of assets' returns. Up until now there was no evidence on the relevance of incorporating current information into the conditional variance modelling in GARCH-type models. I fill this gap by proposing a new model, the RT-GARCH, which incorporates current information. The model is very general; it nests the standard GARCH models as its special case, and can easily incorporate leverage and feedback effects by differentiating between positive and negative news. The new term, i.e. the current realisation of the standardised return, can be viewed in two ways: as a change in the information set, and as an extra shape parameter for the density of returns which determines the "peakedness" and/or thickness of the tails. This shape parameter allows the conditional distribution of returns to have a time-varying kurtosis, which accounting to the empirical application may well play a crucial role in forecasting volatility and VaR during turbulent times.

Estimation of the RT-GARCH revealed that (i) incorporating current information into volatility modelling allows the model to respond quicker to sudden changes of the unconditional level of volatility; and (ii) the combination of ex-ante and ex-post volatility measurement helps to improve out-of-sample volatility forecasts and empirical fit when compared to the forecasts and empirical fit given by the other competing models. Moreover the new model offers a framework for specification testing, which can be thought of a test for constant conditional kurtosis versus a time-varying one.

It would be of interest to investigate whether the empirical performance of the proposed model can be further improved by incorporating some realised measures as in Hansen et al.(2012) and/or assuming Student-t distribution for innovations. In addition, deriving a continuous-time limit of the RT-GARCH model will provide an answer of where exactly between GARCH and SV models it stands. I leave these suggestions for future research.

References

Aït-Sahalia, Y., J. Fan and Y. Li, 2013, The leverage effect puzzle: disentangling sources of bias at high frequency, *Journal of Financial Economics*, 109, 224-249.

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P., 2003, Modeling and forecasting realized volatility, *Econometrica*, 71, 529-626.

Bandi, F. M. and R. Renò, 2012, Time-varying leverage effects, *Journal of Econometrics*, 169, 94-113.

Black, F., 1976, Studies of stock price volatility changes, in *Proceedings of the 1976 Meetings of the American Statistical Association*, 171-181.

Bollerslev, T., 1986, Generalized autoregressive heteroskedasticity, *Journal of Econometrics*, 31, 307-327.

Brandt, A., 1986, The stochastic equation $Y_{n+1} = A_n Y_n + B_n$ with stationary coefficients, *Advances in Applied Probability*, 18, 211-220.

Christie, A. A., 1982, The stochastic behavior of common stock variances: value, leverage and interest rate effects, *Journal of Financial Economics*, 10, 407-432.

Christoffersen, P. F., 1998, Evaluating interval forecasts, *International Economic Review*, 1998, 39(4), 841-862.

Cox, D.R., 1981, Statistical analysis of time series: some recent developments, *Scandinavian Journal of Statistics*, 8(2), 93-115.

Ding, Z., Granger, C.W.J. and Engle, R.F., 1993, A long memory property of stock market returns and a new model, *Journal of Empirical Finance*, 1, 83-106.

Engle, R. F., 1982, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation, *Econometrica*, 50, 987-1007.

Engle, R. F. and V. K. Ng, 1993, Measuring and testing the impact of news on volatility, *The Journal of Finance*, 48, 1749-1778.

Francq, C. and Zakoïan, J.-M., 2004, Maximum likelihood estimation of GARCH and ARMA-GARCH processes, *Bernoulli*, 10, 605-637.

Francq, C. and Zakoïan, J.-M., 2009, Testing the nullity of GARCH coefficients: correction of the standard tests and relative efficiency comparisons, *Journal of the American Statistical Association*, 104, 313-324.

French, K.R., Schwert, G.W. and Stambaugh, R.F., 1987, Expected stock returns and volatility, *Journal of Financial Economics*, 19(1), 3-29.

Ghysels, E., Santa-Clara, P. and Valkanov, R., 2005, There is a risk-return tradeoff after all, *Journal of Financial Economics*, 76, 509-548.

Ghysels, E., Santa-Clara, P. and Valkanov, R., 2006, Predicting volatility: getting the most out of return data sampled at different frequencies, *Journal of Econometrics*, 131, 59-95.

Glosten, L. R., Jagannathan, R. and Runkle, D.E., 1993, On the relation between the expected value and the volatility of the nominal excess return on stocks, *Journal of Finance*, 48(5), 1779-1801.

Hall, P. and Heyde, C., 1980, Martingale limit theory and its application. New York: Academic Press.

Hansen, P.R., 2005, A test for superior predictive ability, *Journal of Business&Economic statistics*, 23(4), 365-380.

Hansen, P.R., Huang, Z. and Shek, H.H., 2011, Realized GARCH: a joint model for returns and realized measures of volatility, *Journal of Applied Econometrics*, 27, 877-906.

Hansen, P.R. and Lunde, A., 2005. A forecast comparison of volatility models: does anything beat a GARCH (1,1)?, *Journal of Applied Econometrics*, 20, 873-889.

Hansen, P.R. and Lunde, A., 2006, Consistent ranking of volatility models, *Journal of Econometrics*, 131, 97-121.

Hansen P.R., Lunde, A. and Nason J. M., 2011, Model confidence set, *Econometrica*, 79, 453-497.

Kesten, H., 1973, Random difference equations and renewal theory for products of random matrices, *Acta Mathematica*, 131(1), 207-248.

Linton, O., Whang Y-J. and Yen, Y-M., 2016, A nonparametric test of a strong leverage hypothesis, *Journal of Econometrics*, 194(1), 153-186.

Meddahi, N. and E. Renault, 2004, Temporal aggregation of volatility models, *Journal of Econometrics*, 119, 355-379.

Nelson, D. B., 1991, Conditional heteroskedasticity in asset returns: a new approach, *Econometrica*, 59, 347-370.

Patton, A., 2011, Volatility forecast evaluation and comparison using imperfect volatility proxies, *Journal of Econometrics*, 160(1), 246-256.

Politis, D. N., 2007, Model-free versus model-based volatility prediction, *Journal of Financial Econometrics*, 5(3), 358-359.

Shephard, N., 2008, Stochastic volatility, in New Palgrave Dictionary of Economics, 2nd Edition, MacMillan.

Shephard, N., 1996, Statistical aspects of ARCH and stochastic volatility, Time Series Models in Econometrics, Finance and Other Fields, London: Chapman & Hall, 1-67.

Smetanina, E., 2018, Supplementary material to “Real-time GARCH”, available at SSRN: http://ssrn.com/abstract=2402617.

Stout, W.F., 1974, Almost sure convergence, probability and mathematical statistics, New York: Academic Press.

Wang, D. C. and P. A. Mykland, 2014, The estimation of leverage effect with high frequency data, *Journal of the American Statistical Association*, 109, 197-215.

Yu, J., 2005, On leverage in a stochastic volatility model, *Journal of Econometrics*, 127, 165-178.

Table 1.1: The conditional variance specification of different models.

RT-GARCH	$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi\epsilon_t^2, \quad E[r_t^2] = E[\lambda_t^2] + 2\varphi$
RT-GARCH with leverage	$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi_1\epsilon_t^2\mathbb{1}_{(\epsilon_t\geq 0)} + \varphi_2\epsilon_t^2\mathbb{1}_{(\epsilon_t<0)}, \quad E[r_t^2] = E[\lambda_t^2] + 2(\varphi_1+\varphi_2)$
RT-GARCH with leverage and feedback	$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma_1 r_{t-1}^2\mathbb{1}_{(r_t\geq 0)} + \gamma_2 r_{t-1}^2\mathbb{1}_{(r_t<0)} + \varphi_1\epsilon_t^2\mathbb{1}_{(\epsilon_t\geq 0)} + \varphi_2\epsilon_t^2\mathbb{1}_{(\epsilon_t<0)}$ $E[r_t^2] = E[\lambda_t^2] + 2(\varphi_1+\varphi_2)$
GARCH(1,1) with standard normal errors	$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma r_{t-1}^2$
GARCH(1,2) with standard normal errors	$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma_1 r_{t-1}^2 + \gamma_2 r_{t-2}^2$
GARCH(1,1) with Student's t-distributed errors	$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma r_{t-1}^2$
GARCH(1,2) with Student's t-distributed errors	$\sigma_t^2 = \alpha + \beta\sigma_{t-1}^2 + \gamma_1 r_{t-1}^2 + \gamma_2 r_{t-2}^2$
APARCH(2,2) with Student's t-distributed errors	$\sigma_t^2 = \alpha_0 + \alpha_1\left[\lvert r_{t-1}\rvert - \gamma_1 r_{t-1}\right]^2 + \alpha_2\left[\lvert r_{t-2}\rvert - \gamma_2 r_{t-2}\right]^2 + \beta_1\sigma_{t-1}^2 + \beta_2\sigma_{t-2}^2$
Simple NoVaS	$\sigma_t^2 = \alpha s_{t-1}^2 + \alpha_0 X_t^2 + \sum_{i=1}^p \alpha_i X_{t-p}^2$ $s_{t-1}^2 = \frac{1}{t-1}\sum_{k=1}^{t-1} X_k^2, \alpha = 0, \alpha_i = \frac{1}{p+1}, 0 \leq i \leq p$
Exponential NoVaS	$\sigma_t^2 = \alpha s_{t-1}^2 + \alpha_0 X_t^2 + \sum_{i=1}^p \alpha_i X_{t-p}^2$ $s_{t-1}^2 = \frac{1}{t-1}\sum_{k=1}^{t-1} X_k^2, \alpha = 0, \alpha_i = c'^{e^{-ci}}, 0 \leq i \leq p, c' = \frac{1}{\sum_{i=0}^p e^{-ci}}$

Note: For Simple NoVaS p is chosen to match the kurtosis (=3) of the normalized return series. For Exponential NoVaS initial value of p is chosen to be $\frac{n}{5}$; c is chosen to match the kurtosis (=3) of the normalized return series and p is adjusted by the maximization routine. APARCH(2,2) corresponds to the standard GJR(2,2) model whenever $0 \leq \gamma_i \leq 1$, $i = 1, 2$.

Table 1.2: Parameter estimates of RT-GARCH models

Parameter estimates of RT-GARCH

Dataset	α	β	γ	φ
IBM	0.0006	0.8755	0.0780	0.0758
	$(18 * 10^{-4})$	$(9 * 10^{-4})$	$(14 * 10^{-4})$	$(21 * 10^{-4})$
GE	0.0001	0.9211	0.0627	0.0378
	$(14 * 10^{-4})$	$(38 * 10^{-3})$	$(2 * 10^{-5})$	$(17 * 10^{-4})$
S&P 500	0.0001	0.9124	0.0726	0.0138
	$(12 * 10^{-4})$	$(14 * 10^{-3})$	$(45 * 10^{-3})$	$(11 * 10^{-4})$

Parameter estimates of RT-GARCH with leverage

Dataset	α	β	γ	φ_1	φ_2
IBM	0.0003	0.8883	0.0703	0.0475	0.0886
	$(15 * 10^{-4})$	$(6 * 10^{-4})$	$(11 * 10^{-4})$	$(19 * 10^{-4})$	$(27 * 10^{-4})$
GE	0.0001	0.9273	0.0550	0.0237	0.0529
	$(2.7 * 10^{-4})$	$(38 * 10^{-4})$	$(4.2 * 10^{-4})$	$(2 * 10^{-4})$	$(48 * 10^{-3})$
S&P 500	0.0016	0.8995	0.0718	0.0003	0.0481
	$(25 * 10^{-4})$	$(15 * 10^{-3})$	$(6.7 * 10^{-4})$	$(27 * 10^{-4})$	$(8.1 * 10^{-4})$

Parameter estimates of RT-GARCH with leverage and feedback

Dataset	α	β	γ_1	γ_2	φ_1	φ_2
IBM	0.0001	0.8599	0.0328	0.0706	0.0903	0.1319
	$(17 * 10^{-4})$	$(30 * 10^{-4})$	$(14 * 10^{-4})$	$(15 * 10^{-4})$	$(26 * 10^{-4})$	$(29 * 10^{-4})$
GE	0.0001	0.9225	0.0343	0.0322	0.0450	0.1253
	$(15 * 10^{-4})$	$(40 * 10^{-3})$	$(11 * 10^{-4})$	$(10 * 10^{-4})$	$(18 * 10^{-3})$	$(27 * 10^{-3})$
S&P 500	0.0023	0.9185	0.0127	0.0605	0.0004	0.0740
	$(4.2 * 10^{-4})$	$(1.8 * 10^{-3})$	$(5 * 10^{-4})$	$(23 * 10^{-4})$	(10^{-4})	$(4.6 * 10^{-4})$

Note: The table presents parameter estimates for respective models based on the full sample. The sample size used for estimation is 3000 for IBM and GE stocks and 2000 for SP500 index. Standard errors, calculated numerically, are given in parentheses.

Table 1.3: Forecast evaluation based on MSE loss (full sample)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	e p_{MCS}	MSE	p_{MCS}
RT-GARCH	5.9626	0.0940⋆	6.7355	0.2530⋆	2.5227	0.0400
RT-GARCH-L	5.8989	0.0990⋆	6.6591	0.9170⋆	2.0289	0.6220⋆
RT-GARCH-LF	6.0861	0.0820⋆	6.6274	1⋆	1.9370	1⋆
A-PARCH(2,2)-$St.t$ distr.	5.8152	0.6330⋆	6.6632	0.9170⋆	2.6657	0.0020
GARCH(1,1)-$N(0,1)$	5.9074	0.0990⋆	6.8069	0.0070	2.3780	0.0450
GARCH(1,2)-$N(0,1)$	5.9074	0.0990⋆	6.8069	0.0070	2.3780	0.0450
GARCH(1,1)-$St.t$ distr.	5.7074	0.7470⋆	6.8199	0.0030	2.3725	0.0450
GARCH(1,2)-$St.t$ distr.	5.6923	1⋆	6.9611	0.0030	2.5405	0.0310
Simple NoVaS	7.9097	0	8.6489	0	2.6415	0.0060
Exponential NoVaS	7.9305	0	8.7922	0	2.7714	0.0010

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	5.4655	0.0660⋆	6.1762	0.1130⋆	1.7490	0.1820⋆
RT-GARCH-L	5.7604	0.0010	6.1410	0.1130⋆	1.6724	1⋆
RT-GARCH-LF	7.4039	0.0005	6.8305	0.0005	2.7361	0.0600⋆
A-PARCH(2,2)-$St.t$ distr.	5.5588	0.0200	6.2182	0.1130⋆	2.0496	0.1230⋆
GARCH(1,1)-$N(0,1)$	5.5436	0.0200	6.2221	0.0440	1.7265	0.1820⋆
GARCH(1,2)-$N(0,1)$	5.5436	0.0020	6.2221	0.0440	1.7265	0.1820⋆
GARCH(1,1)-$St.t$ distr.	5.2380	0.1460⋆	6.2346	0.0220	2.1047	0.1090⋆
GARCH(1,2)-$St.t$ distr.	5.1158	1⋆	5.9126	1⋆	2.1059	0.1090⋆
Simple NoVaS	7.9295	0	8.8844	0	2.8555	0.0220
Exponential NoVaS	7.9810	0	8.9874	0	2.9460	0.0050

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.4: Forecast evaluation based on MSE loss (full sample)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	5.5838	0.6440*	6.5120	0.5320*	1.8604	1*
RT-GARCH-L	6.7789	0.0020	6.5751	0.0570*	2.3635	0.0020
RT-GARCH-LF	7.1880	0	8.1607	0	2.4339	0.0220
A-PARCH(2,2)-$St.t$ distr.	5.2557	1*	6.3874	0.8660*	2.1500	0.1010*
GARCH(1,1)-$N(0,1)$	5.9061	0.0410	6.7387	0.0110	2.1701	0.0380
GARCH(1,2)-$N(0,1)$	5.9061	0.0410	6.7387	0.0110	2.1701	0.0380
GARCH(1,1)-$St.t$ distr.	5.7090	0.0930*	6.7123	0.0150	2.2999	0.0130
GARCH(1,2)-$St.t$ distr.	5.7043	0.0930*	6.3749	1*	2.4080	0.0020
Simple NoVaS	7.4780	0	9.0686	0	3.0637	0
Exponential NoVaS	7.4882	0	9.1016	0	3.0515	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	6.2789	0.1290*	6.8308	0.2120*	2.0271	1*
RT-GARCH-L	7.3068	0.0030	7.1004	0.0030	2.8156	0.0070
RT-GARCH-LF	7.7454	0.0010	9.5520	0.0030	2.8458	0.0010
A-PARCH(2,2)-$St.t$ distr.	5.6200	1*	6.5759	1*	2.3697	0.0740*
GARCH(1,1)-$N(0,1)$	6.4643	0.0130	7.4280	0.0020	2.5460	0.0220
GARCH(1,2)-$N(0,1)$	6.4643	0.0130	7.4280	0.0020	2.5460	0.0220
GARCH(1,1)-$St.t$ distr.	7.0099	0.0020	7.3615	0.0020	2.5960	0.0240
GARCH(1,2)-$St.t$ distr.	5.7186	0.7550*	6.8867	0.0290	2.7077	0.0090
Simple NoVaS	8.0218	0	9.1868	0	3.2885	0
Exponential NoVaS	8.0104	0	9.2016	0	3.1058	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.5: Forecast evaluation based on QLIKE loss (full sample)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4628	0.5500*	1.4505	0.0610*	0.9426	0.3740*
RT-GARCH-L	1.4531	1*	1.4315	1*	0.9471	0.1550*
RT-GARCH-LF	1.4828	0.3800*	1.4384	0.6540*	1.0077	0.0010
A-PARCH(2,2)-$St.t$ distr.	1.5487	0.0120	1.4733	0.0320	0.9322	1*
GARCH(1,1)-$N(0,1)$	1.5106	0.0320	1.4767	0.0270	0.9488	0.0450
GARCH(1,2)-$N(0,1)$	1.5106	0.0320	1.4767	0.0270	0.9488	0.0450
GARCH(1,1)-$St.t$ distr.	1.5252	0.0400	1.4786	0.0180	0.9494	0.0440
GARCH(1,2)-$St.t$ distr.	1.5277	0.0400	1.4786	0.0180	0.9376	0.4540*
Simple NoVaS	3.9372	0	3.3669	0	1.4625	0.0005
Exponential NoVaS	3.9361	0	3.3633	0	1.5563	0.0005

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.3935	1*	1.3760	1*	0.8957	0.2950*
RT-GARCH-L	1.4834	0.0140	1.3932	0.0560*	1.0169	0.0090
RT-GARCH-LF	1.6347	0.0040	1.5198	0.0010	1.2279	0.0040
A-PARCH(2,2)-$St.t$ distr.	1.4492	0.0170	1.4034	0.0560*	0.8955	1*
GARCH(1,1)-$N(0,1)$	1.4157	0.0280	1.4163	0.0110	0.9159	0.0310
GARCH(1,2)-$N(0,1)$	1.4157	0.0280	1.4163	0.0110	0.9159	0.0310
GARCH(1,1)-$St.t$ distr.	1.4131	0.0300	1.3948	0.0560	0.9470	0.0120
GARCH(1,2)-$St.t$ distr.	1.4144	0.0300	1.3987	0.0560	0.9425	0.0120
Simple NoVaS	4.1445	0	3.6022	0	1.6679	0.0020
Exponential NoVaS	4.0942	0	3.5691	0	1.8015	0.0020

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

Table 1.6: Forecast evaluation based on QLIKE loss (full sample)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4462	0.3600*	1.3985	1*	0.9106	0.6920*
RT-GARCH-L	1.5924	0.0220	1.4476	0.0520	1.1241	0.0090
RT-GARCH-LF	1.8417	0.0130	1.6762	0.0010	1.4544	0.0010
A-PARCH(2,2)-$St.t$ distr.	1.4177	1*	1.4050	0.8220*	0.9081	1*
GARCH(1,1)-$N(0,1)$	1.4864	0.0490	1.4736	0.0010	0.9734	0.0200
GARCH(1,2)-$N(0,1)$	1.4864	0.0490	1.4736	0.0010	0.9734	0.0200
GARCH(1,1)-$St.t$ distr.	1.5101	0.0330	1.4684	0.0010	1.0058	0.0200
GARCH(1,2)-$St.t$ distr.	1.5088	0.0330	1.4014	0.0010	0.9620	0.0380
Simple NoVaS	3.2675	0	3.8264	0	1.9397	0.0005
Exponential NoVaS	3.2776	0	3.7663	0	2.0835	0.0005

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4995	0.1890*	1.4267	0.2660*	0.9396	0.5410*
RT-GARCH-L	1.6823	0.0180	1.5039	0.0010	1.1974	0.0080
RT-GARCH-LF	1.7036	0.0090	1.8094	0.0005	1.3169	0.0010
A-PARCH(2,2)-$St.t$ distr.	1.4279	1*	1.4122	1*	0.9352	1*
GARCH(1,1)-$N(0,1)$	1.5533	0.0420	1.5425	0.0050	1.0357	0.0350
GARCH(1,2)-$N(0,1)$	1.5533	0.0420	1.5425	0.0050	1.0357	0.0350
GARCH(1,1)-$St.t$ distr.	1.6097	0.0270	1.5354	0.0020	1.0610	0.0110
GARCH(1,2)-$St.t$ distr.	1.5063	0.0270	1.4534	0.0410	1.0129	0.0700*
Simple NoVaS	3.5635	0	3.8891	0	2.1970	0
Exponential NoVaS	3.5632	0	3.8002	0	2.3442	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.7: Forecast evaluation based on MSE loss (pre-crisis period)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	2.2499	0.8880*	2.9478	0.5310*	1.1053	0.3540*
RT-GARCH-L	2.2334	0.8880*	2.9211	0.5310*	1.1198	0.1810*
RT-GARCH-LF	2.3112	0.3370*	2.9234	0.5310*	0.8193	1*
A-PARCH(2,2)-$St.t$ distr.	2.1908	1*	2.8280	1*	0.8508	0.7260*
GARCH(1,1)-$N(0,1)$	2.3644	0.3000*	3.0881	0.0850*	1.1257	0.0470
GARCH(1,2)-$N(0,1)$	2.3644	0.3000*	3.0881	0.0850*	1.1358	0.0470
GARCH(1,1)-$St.t$ distr.	2.2873	0.7700*	3.1117	0.0690*	1.1677	0.0423
GARCH(1,2)-$St.t$ distr.	2.2930	0.7700*	3.1117	0.0690*	1.1716	0.0410
Simple NoVaS	3.5073	0.0010	4.8216	0.0010	1.4206	0.0100
Exponential NoVaS	3.5755	0.0010	4.4485	0.0010	1.4428	0.0100

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	1.9340	0.7160*	2.4904	0.5893*	1.0603	1*
RT-GARCH-L	1.6637	1*	2.5724	0.1220*	1.2708	0.0450
RT-GARCH-LF	2.1221	0.1461*	2.8155	0.0430	1.1718	0.1150*
A-PARCH(2,2)-$St.t$ distr.	2.0441	0.3700*	2.2552	1*	1.1618	0.1150*
GARCH(1,1)-$N(0,1)$	2.0740	0.1460*	2.8630	0.0290	1.2890	0.0410
GARCH(1,2)-$N(0,1)$	2.0740	0.1460*	2.8630	0.0290	1.2909	0.0410
GARCH(1,1)-$St.t$ distr.	2.0349	0.2400*	2.8896	0.0220	1.1869	0.1950*
GARCH(1,2)-$St.t$ distr.	2.0311	0.2421*	2.8896	0.0220	1.1300	0.1950*
Simple NoVaS	3.7046	0	4.7204	0.0080	1.5833	0.0150
Exponential NoVaS	3.6552	0	4.8017	0.0090	1.5126	0.0150

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.8: Forecast evaluation based on MSE loss (pre-crisis period)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	2.0836	0.1470⋆	2.3709	0.0739⋆	1.2375	0.6940⋆
RT-GARCH-L	1.9729	1⋆	2.2761	1⋆	1.1472	1⋆
RT-GARCH-LF	2.3842	0.1000⋆	2.4838	0.0219	1.4586	0.0415
A-PARCH(2,2)-$St.t$ distr.	2.2134	0.1370⋆	2.4967	0.0219	1.3357	0.4150⋆
GARCH(1,1)-$N(0,1)$	2.0265	0.3590⋆	2.7944	0.0180	1.3771	0.0381
GARCH(1,2)-$N(0,1)$	2.0265	0.3590⋆	2.7044	0.0180	1.3788	0.0255
GARCH(1,1)-$St.t$ distr.	2.0823	0.1470⋆	2.8173	0.0150	1.3468	0.2550⋆
GARCH(1,2)-$St.t$ distr.	2.0797	0.1470⋆	2.8173	0.0150	1.3479	0.2410⋆
Simple NoVaS	3.9311	0	5.0234	0	1.6486	0.0380
Exponential NoVaS	3.9594	0	5.0399	0	1.4911	0.0380

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	2.0443	1⋆	2.4923	0.7170⋆	1.3236	0.7651⋆
RT-GARCH-L	2.3620	0.0130	2.4617	1⋆	1.3178	1⋆
RT-GARCH-LF	2.1540	0.0510⋆	2.7285	0.1210⋆	1.4642	0.0342
A-PARCH(2,2)-$St.t$ distr.	2.1910	0.0460	2.7849	0.0650⋆	1.3413	0.3860⋆
GARCH(1,1)-$N(0,1)$	2.2150	0.0160	2.8142	0.0550⋆	1.4328	0.0120
GARCH(1,2)-$N(0,1)$	2.2150	0.0160	2.8142	0.0550⋆	1.4842	0.0120
GARCH(1,1)-$St.t$ distr.	2.1409	0.1090⋆	2.8321	0.0507⋆	1.3931	0.3420⋆
GARCH(1,2)-$St.t$ distr.	2.1398	0.1220⋆	2.8321	0.0507⋆	1.3994	0.3420⋆
Simple NoVaS	4.2298	0.0060	5.1980	0.0050	1.6308	0.0040
Exponential NoVaS	4.1846	0.0060	5.0943	0.0130	1.4760	0.0040

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.9: Forecast evaluation based on QLIKE loss (pre-crisis period)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.6258	0.0250	1.6284	0.1046*	1.2023	0.0580*
RT-GARCH-L	1.5958	0.9310*	1.6277	0.2380*	1.1762	0.0610*
RT-GARCH-LF	1.6602	0.0403	1.6272	0.4090*	1.1398	1*
A-PARCH(2,2)-$St.t$ distr.	1.5946	1*	1.6074	1*	1.1509	0.0750*
GARCH(1,1)-$N(0,1)$	1.6114	0.2401*	1.6591	0.0126	1.2110	0.0180
GARCH(1,2)-$N(0,1)$	1.6114	0.2401*	1.6591	0.0126	1.2043	0.0480
GARCH(1,1)-$St.t$ distr.	1.6064	0.2870*	1.6643	0.0140	1.1925	0.0350
GARCH(1,2)-$St.t$ distr.	1.6051	0.2870*	1.6643	0.0140	1.2084	0.0410
Simple NoVaS	2.6016	0.0077	2.6602	0	1.7735	0.0050
Exponential NoVaS	2.5376	0.0077	2.6812	0	1.7158	0.0010

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.5482	0.7720*	1.5588	0.5490*	1.2017	1*
RT-GARCH-L	1.5433	1*	1.5722	0.0910*	1.2427	0.1080*
RT-GARCH-LF	1.6091	0.0559*	1.5973	0.0420	1.2035	0.5640*
A-PARCH(2,2)-$St.t$ distr.	1.5782	0.0853*	1.5468	1*	1.2340	0.2820*
GARCH(1,1)-$N(0,1)$	1.5558	0.4340*	1.6172	0.0160	1.2687	0.0408
GARCH(1,2)-$N(0,1)$	1.5558	0.4340*	1.6172	0.0160	1.2687	0.0408
GARCH(1,1)-$St.t$ distr.	1.5660	0.1550*	1.6121	0.0220	1.2559	0.0440
GARCH(1,2)-$St.t$ distr.	1.5656	0.1550*	1.6121	0.0220	1.2559	0.0440
Simple NoVaS	2.7961	0	3.0279	0	1.9897	0.0130
Exponential NoVaS	2.6950	0	2.9584	0.0030	1.8741	0.0160

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

Table 1.10: Forecast evaluation based on QLIKE loss (pre-crisis period)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.5591	0.0920⋆	1.5399	1⋆	1.2256	1⋆
RT-GARCH-L	1.5499	1⋆	1.5493	0.2980⋆	1.2311	0.5440⋆
RT-GARCH-LF	1.5598	0.0520⋆	1.5802	0.0650⋆	1.2767	0.0453
A-PARCH(2,2)-$St.t$ distr.	1.5990	0.0213	1.6078	0.0060	1.2396	0.1238⋆
GARCH(1,1)-$N(0,1)$	1.5630	0.0260	1.5917	0.0390	1.3013	0.0238
GARCH(1,2)-$N(0,1)$	1.5630	0.0260	1.5917	0.0390	1.2985	0.0161
GARCH(1,1)-$St.t$ distr.	1.5694	0.0201	1.5957	0.0100	1.2716	0.0482⋆
GARCH(1,2)-$St.t$ distr.	1.5691	0.0201	1.5957	0.0100	1.2766	0.0490⋆
Simple NoVaS	3.0539	0	3.3247	0	2.0528	0.0130
Exponential NoVaS	2.8865	0	3.1516	0	1.9589	0.0260

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.5572	0.2800⋆	1.5547	0.3380⋆	1.2486	1⋆
RT-GARCH-L	1.5473	0.7015⋆	1.5474	1⋆	1.2558	0.4810⋆
RT-GARCH-LF	1.5339	1⋆	1.6121	0.0290	1.3107	0.0449
A-PARCH(2,2)-$St.t$ distr.	1.6039	0.0352	1.5942	0.0390	1.2982	0.1565⋆
GARCH(1,1)-$N(0,1)$	1.5672	0.0390	1.5852	0.0450	1.3129	0.0419
GARCH(1,2)-$N(0,1)$	1.5672	0.0390	1.5852	0.0450	1.3105	0.0449
GARCH(1,1)-$St.t$ distr.	1.5695	0.0352	1.5884	0.0420	1.3107	0.0440
GARCH(1,2)-$St.t$ distr.	1.5693	0.0352	1.5884	0.0420	1.3107	0.0440
Simple NoVaS	3.3053	0	3.4334	0.0020	2.0618	0.0040
Exponential NoVaS	3.1063	0.0030	3.1453	0.0010	2.0030	0.0031

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.11: Forecast evaluation based on MSE loss (crisis and post-crisis period)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	4.5716	0.0020	11.4981	0.3600*	1.8311	0.0350
RT-GARCH-L	4.4319	0.6390*	11.3606	1*	1.6040	0.4080*
RT-GARCH-LF	4.4203	1*	11.6326	0.0020	1.3981	1*
A-PARCH(2,2)-$St.t$ distr.	4.4613	0.3390*	11.5223	0.0415	2.3010	0
GARCH(1,1)-$N(0,1)$	4.4750	0.0300	11.6015	0.0050	1.7184	0.0470
GARCH(1,2)-$N(0,1)$	4.4750	0.0300	11.6015	0.0050	1.7184	0.0470
GARCH(1,1)-$St.t$ distr.	4.4878	0.0250	11.5367	0.0400	1.9842	0.0020
GARCH(1,2)-$St.t$ distr.	4.6584	0.0020	11.6527	0.0010	2.4677	0
Simple NoVaS	5.5788	0	12.0917	0	1.9697	0.0020
Exponential NoVaS	5.6237	0	12.1720	0	1.9331	0.0080

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	4.1624	0.0960*	10.4028	0.2690*	0.8056	1*
RT-GARCH-L	4.1081	0.7230*	10.2550	1*	0.9355	0.0810*
RT-GARCH-LF	4.2980	0.0960*	11.2363	0	1.6040	0.0010
A-PARCH(2,2)-$St.t$ distr.	3.9632	1*	10.6841	0.0110	1.3555	0.0020
GARCH(1,1)-$N(0,1)$	4.2318	0.0310	10.7131	0.0110	1.0701	0.0110
GARCH(1,2)-$N(0,1)$	4.2318	0.0310	10.7131	0.0110	1.0701	0.0110
GARCH(1,1)-$St.t$ distr.	4.2136	0.0400	10.5373	0.0410	1.4371	0.0030
GARCH(1,2)-$St.t$ distr.	4.2136	0.0400	10.5325	0.0410	1.3019	0.0030
Simple NoVaS	5.7350	0	13.3463	0	1.8017	0
Exponential NoVaS	5.7397	0	13.3697	0	1.7878	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

Table 1.12: Forecast evaluation based on MSE loss (crisis and post-crisis period)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	4.6090	0.0660*	10.9616	1*	1.1098	1*
RT-GARCH-L	4.7146	0.0610*	11.1144	0.5020*	1.3585	0.1250*
RT-GARCH-LF	4.9129	0.0270	11.2357	0.5020*	2.9225	0.0005
A-PARCH(2,2)-$St.t$ distr.	4.4016	1*	11.1299	0.5020*	1.7746	0.0150
GARCH(1,1)-$N(0,1)$	4.6784	0.0420	11.3812	0.0030	1.5144	0.0250
GARCH(1,2)-$N(0,1)$	4.6784	0.0420	11.3812	0.0030	1.5144	0.0250
GARCH(1,1)-$St.t$ distr.	4.6889	0.0420	11.3346	0.0050	2.5789	0.0005
GARCH(1,2)-$St.t$ distr.	4.6519	0.0420	11.3889	0.0030	2.7429	0.0005
Simple NoVaS	3.9311	0	13.5452	0	2.9198	0.0005
Exponential NoVaS	3.9594	0	13.5259	0	2.8922	0.0005

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	MSE	p_{MCS}	MSE	p_{MCS}	MSE	p_{MCS}
RT-GARCH	4.7423	0.6190*	11.3676	1*	1.7102	1*
RT-GARCH-L	5.0036	0.2210*	11.8478	0.0130	2.1316	0.1100*
RT-GARCH-LF	5.7171	0.0200	11.4883	0.5870*	3.6060	0.0190
A-PARCH(2,2)-$St.t$ distr.	4.6119	1*	11.4266	0.5870*	2.8358	0.0190
GARCH(1,1)-$N(0,1)$	5.7232	0.0200	11.6886	0.0350	2.1965	0.0410
GARCH(1,2)-$N(0,1)$	5.7232	0.0200	11.6886	0.0350	2.1965	0.0410
GARCH(1,1)-$St.t$ distr.	5.6944	0.0200	11.8160	0.0150	4.3403	0
GARCH(1,2)-$St.t$ distr.	5.7463	0.0200	11.7441	0.0150	4.0890	0
Simple NoVaS	5.9670	0.0100	13.6172	0	3.3435	0.0190
Exponential NoVaS	5.9310	0.0100	13.6538	0	3.3017	0.0190

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.13: Forecast evaluation based on QLIKE loss (crisis and post-crisis period)

1-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4640	0.0400	1.4299	0.0680*	0.6870	0.2500*
RT-GARCH-L	1.4069	0.5150*	1.3717	1*	0.7022	0.0340
RT-GARCH-LF	1.4054	1*	1.4097	0.3960*	0.7582	0.0340
A-PARCH(2,2)-$St.t$ distr.	1.4403	0.0640*	1.4245	0.0680*	0.6751	1*
GARCH(1,1)-$N(0,1)$	1.4533	0.0420	1.4573	0.0470	0.7815	0.0150
GARCH(1,2)-$N(0,1)$	1.4533	0.0420	1.4573	0.0470	0.7815	0.0150
GARCH(1,1)-$St.t$ distr.	1.4577	0.0420	1.4410	0.0470	0.6912	0.0340
GARCH(1,2)-$St.t$ distr.	1.5020	0.0380	1.4783	0.0090	0.6980	0.0340
Simple NoVaS	3.4775	0	6.0813	0	1.2339	0
Exponential NoVaS	3.5190	0	6.1866	0	1.3122	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

5-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.3452	0.8750*	1.2648	1*	0.5482	1*
RT-GARCH-L	1.3743	0.0110	1.2444	0.3790*	0.6770	0.0200
RT-GARCH-LF	1.4143	0	1.3115	0.0020	0.8759	0.0010
A-PARCH(2,2)-$St.t$ distr.	1.3205	1*	1.2974	0.0020	0.5514	0.7820*
GARCH(1,1)-$N(0,1)$	1.3635	0.0110	1.2792	0.0370	0.6334	0.0340
GARCH(1,2)-$N(0,1)$	1.3635	0.0110	1.2792	0.0370	0.6334	0.0340
GARCH(1,1)-$St.t$ distr.	1.3552	0.0160	1.2861	0.0020	0.7137	0.0010
GARCH(1,2)-$St.t$ distr.	1.3552	0.0160	1.2794	0.0370	0.5869	0.0440
Simple NoVaS	3.9533	0	6.4805	0	1.3514	0
Exponential NoVaS	3.8880	0	6.6783	0	1.4072	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^\star_{95\%}$.

Table 1.14: Forecast evaluation based on QLIKE loss (crisis and post-crisis period)

10-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4416	0.0580*	1.2917	1*	0.5835	1*
RT-GARCH-L	1.4579	0.0070	1.3534	0.0010	0.7893	0.0030
RT-GARCH-LF	1.5612	0	1.3280	0.0870*	1.0837	0
A-PARCH(2,2)-*St.t* distr.	1.3864	1*	1.3162	0.0870*	0.5922	0.5530*
GARCH(1,1)-$N(0,1)$	1.4577	0.0050	1.3311	0.0030	0.7282	0.0030
GARCH(1,2)-$N(0,1)$	1.4577	0.0050	1.3311	0.0030	0.7282	0.0030
GARCH(1,1)-*St.t* distr.	1.4336	0.0630*	1.3483	0.0020	0.9085	0.0010
GARCH(1,2)-*St.t* distr.	1.4399	0.0530*	1.3232	0.0030	0.7230	0.0030
Simple NoVaS	3.7833	0	7.5707	0	1.5315	0
Exponential NoVaS	3.8311	0	6.8522	0	1.5865	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

15-step ahead volatility forecasts

	IBM		**GE**		**S&P 500**	
Model	QLIKE	p_{MCS}	QLIKE	p_{MCS}	QLIKE	p_{MCS}
RT-GARCH	1.4751	0.1410*	1.3163	1*	0.7297	1*
RT-GARCH-L	1.4933	0.1410*	1.4231	0.0030	0.9511	0.0080
RT-GARCH-LF	1.6641	0.0160	1.3511	0.0810*	1.2957	0.0080
A-PARCH(2,2)-*St.t* distr.	1.4296	1*	1.3311	0.2290*	0.7496	0.1790*
GARCH(1,1)-$N(0,1)$	1.5646	0.0200	1.3736	0.0160	0.8839	0.0080
GARCH(1,2)-$N(0,1)$	1.5646	0.0200	1.3736	0.0160	0.8839	0.0080
GARCH(1,1)-*St.t* distr.	1.4537	0.2020*	1.3980	0.0040	1.1265	0.0080
GARCH(1,2)-*St.t* distr.	1.4616	0.1410*	1.3613	0.0300	1.1065	0.0080
Simple NoVaS	3.3053	0	6.7079	0	1.8328	0
Exponential NoVaS	3.1063	0.0030	6.9641	0	1.8260	0

Note: p_{MCS} are the p-values from Model Confidence Set test of Hansen et al.(2011). The p-values that are marked with a $\star$ are those in the model confidence set $\widehat{\mathcal{M}}^{\star}_{95\%}$.

Table 1.15: Evaluation of 1-step ahead VaR(5%) forecasts (full sample).

	IBM		GE		S&P 500	
Model	VR	LR_{cc}	VR	LR_{cc}	VR	LR_{cc}
RT-GARCH	1.08	5.3119	0.8000	5.1842	0.6700	5.6921
RT-GARCH-L	0.9800	5.5283	0.6400	5.4509	0.5800	5.7488
RT-GARCH-LF	0.3600	5.8141	0.4600	5.4396	0.3800	5.7207
A-PARCH(2,2)-$St.t$ distr.	0.7800	5.4278	0.2800	5.7620	0.2000	5.8207
GARCH(1,1)-$N(0,1)$	0.7800	5.4278	0.2200	5.8057	0.2000	5.8207
GARCH(1,2)-$N(0,1)$	0.7600	5.4995	0.3800	5.6248	0.3400	5.7202
GARCH(1,1)-$St.t$ distr.	0.3600	5.8141	0.3000	5.7478	0.1400	5.8668
GARCH(1,2)-$St.t$ distr.	0.5400	5.6659	0.3200	5.7336	0.1800	5.8390

Note: VR denotes the violation ratio=(# of returns that exceed the VaR(5%) forecast)/(# of the expected violations); LR_{cc} stands for Likelihood Ratio test for conditional coverage, see Christoffersen (1998). Moreover, $LR_{cc} \sim \chi^2_{(2)}$ with critical values 5.99 ($p = 0.05$) and 9.21 ($p = 0.01$).

1.9 Appendix A.

.

To derive the eq.(1.5) observe that λ_t^2 cn be written as follows:

$$\lambda_t^2 = b_{t-1} + \varphi\epsilon_t^2 = b_{t-1} + \varphi\frac{r_t^2}{\lambda_t^2}.$$

Provided that $\lambda_t^2 > 0$, it follows that:

$$\lambda_t^2 = \frac{1}{2}b_{t-1}+\frac{1}{2}\sqrt{b_{t-1}^2+4\varphi r_t^2} = \frac{1}{2}b_{t-1}+\frac{1}{2}b_{t-1}\sqrt{1+\frac{4\varphi r_t^2}{b_{t-1}^2}} = \frac{1}{2}b_{t-1}+\frac{1}{2}b_{t-1}\left(1+\frac{2\varphi r_t^2}{b_{t-1}^2}\right)+o\left(\frac{r_t^2}{b_{t-1}}\right) =$$
$$= b_{t-1}+\frac{\varphi r_t^2}{b_{t-1}}+o\left(\frac{r_t^2}{b_{t-1}}\right) = \frac{\varphi r_t^2}{b_{t-1}}+o\left(\frac{r_t^2}{b_{t-1}}\right)+\alpha+\gamma r_{t-1}^2+\beta\lambda_{t-1}^2 = \frac{\varphi r_t^2}{b_{t-1}}+o\left(\frac{r_t^2}{b_{t-1}}\right)+\alpha+\gamma r_{t-1}^2+$$
$$+\beta\left[\frac{\varphi r_{t-1}^2}{b_{t-2}}+\alpha+\gamma r_{t-2}^2+\beta\lambda_{t-2}^2+o\left(\frac{r_{t-1}^2}{b_{t-2}}\right)\right] = \cdots =$$
$$= \frac{\alpha}{1-\beta}+\frac{\varphi r_t^2}{b_{t-1}}+\sum_{j=1}^{\infty}\left(\frac{\beta^j\varphi}{b_{t-1-j}}+\gamma\beta^{j-1}\right)r_{t-j}^2+\sum_{j=0}^{\infty}o\left(\frac{r_{t-j}^2}{b_{t-1-j}}\right),$$

where in the first line of the derivations I used that for $x << 1$ it holds that $(1+x)^\alpha \approx 1+\alpha x+o(x)$.

Proof of Theorem 1. Consider the general model:

$$r_t = \lambda_t\epsilon_t$$

$$\lambda_t^2 = \alpha+\beta\lambda_{t-1}^2+\gamma r_{t-1}^2+\varphi\epsilon_t^2,$$

where $\{\epsilon_t\}$ are i.i.d. random variables such that $E\left[\epsilon_t\right] = 0$, $E\left[\epsilon_t^2\right] = 1$ with the density f_ϵ. Denote by $P\left(r_t \le c|\mathcal{F}_{t-1}\right)$ the conditional cumulative probability function, where $\mathcal{F}_{t-1}$ denotes the information set up to time $t-1$. In order to compute $P(r_t \le c|\mathcal{F}_{t-1})$ note that the first equation can be rewritten as

$$r_t = \sqrt{\alpha+\beta\lambda_{t-1}^2+\gamma r_{t-1}^2+\varphi\epsilon_t^2}\epsilon_t$$

such that

$$P\left(r_t \le c|\mathcal{F}_{t-1}\right) = P(\sqrt{\alpha+\beta\lambda_{t-1}^2+\gamma r_{t-1}^2+\varphi\epsilon_t^2}\epsilon_t \le c|\mathcal{F}_{t-1}).$$

Let e be any realisation of ϵ_t, such that conditional on $\mathcal{F}_{t-1}$ the following condition holds:

$$\sqrt{\alpha+\beta\lambda_{t-1}^2+\gamma r_{t-1}^2+\varphi e^2}e \le c$$

Define d to be the largest value of e. To obtain d I first square the above equation such that

$$(\alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi d^2)d^2 = c^2 \Leftrightarrow (\alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2)d^2 + \varphi d^4 = c^2. \tag{1.22}$$

Eq. (1.22) is a quartic equation in d whenever $\varphi \neq 0$ and is quadratic equation in d whenever $\varphi = 0$ (which is simply the usual GARCH(1,1) case). For quartic equation the solutions are given by:

$$d_{1,2} = \pm\sqrt{\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} - b_{t-1}}{2\varphi}}$$

$$d_{3,4} = \pm\sqrt{-\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} + b_{t-1}}{2\varphi}}$$

with $b_{t-1} = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2$. I disregard $d_{3,4}$ since I am only interested in the real valued solutions, such that I have:

$$d(c) = sign(c)\sqrt{\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} - b_{t-1}}{2\varphi}} \tag{1.23}$$

and

$$P\left(r_t \leq c|\mathcal{F}_{t-1}\right) = \int_{-\infty}^{d(c)} f_\epsilon(x)dx.$$

In order to emphasise the dependence of $d(c)$ on the past information as well as the parameter vector $\theta = (\alpha, \beta, \gamma, \varphi)'$ I adopt the following notation:

$$d(c, b_{t-1}, \theta) = sign(c)\sqrt{\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} - b_{t-1}}{2\varphi}}. \tag{1.24}$$

The solution to the quadratic equation in d for the case $\varphi = 0$ is given by:

$$d(c, b_{t-1}, \theta) = c/\sqrt{b_{t-1}}, \tag{1.25}$$

which corresponds to the standard GARCH(1,1) model (as $b_{t-1} = \sigma_t^2$ whenever $\varphi = 0$), for which the conditional density of the returns is just $f_r(r|\mathcal{F}_{t-1}) = \frac{1}{\sqrt{b_{t-1}}} f_\epsilon(\epsilon) = f_\epsilon(\epsilon)/\sigma_t$. To obtain the density in the case $\varphi \neq 0$ and $r \neq 0$ I use Leibniz integral rule with variable limits to get:

$$f_r(r|\mathcal{F}_{t-1}) = \frac{\partial P(r_t \leq c|\mathcal{F}_{t-1})}{\partial c}\,|_{c=r} = \frac{\partial d(c, b_{t-1}, \theta)}{\partial c}\,|_{c=r}\ f_\epsilon(d(r, b_{t-1}, \theta)) =$$

$$\left\{\frac{\partial sign(c)}{\partial c}\sqrt{\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} - b_{t-1}}{2\varphi}} + sign(c)\frac{1}{2}\left(\frac{\sqrt{b_{t-1}^2 + 4c^2\varphi} - b_{t-1}}{2\varphi}\right)^{-\frac{1}{2}}\frac{1}{2\varphi}\times\right.$$

$$\left.\times\,\frac{1}{2}(b_{t-1}^2 + 4c^2\varphi)^{-\frac{1}{2}}8c\varphi\right\}_{c=r} f_\epsilon(d(r, b_{t-1}, \theta)) =$$

$$= sign(r)r\sqrt{\frac{2\varphi}{(b_{t-1}^2 + 4r^2\varphi)(\sqrt{b_{t-1}^2 + 4r^2\varphi} - b_{t-1})}}f_\epsilon(d(r, b_{t-1}, \theta)) =$$

$$= \frac{|r|}{d(r, b_{t-1}, \theta)\sqrt{b_{t-1}^2 + 4r^2\varphi}}f_\epsilon(d(r, b_{t-1}, \theta)).$$

Before I calculate the limit of the above equation at $r = 0$, note that $d(r, b_{t-1}, \theta)$ in the denominator involves $sign(r)$, while the numerator involves $|r| = rsign(r)$, I thus can write the density as:

$$f_r(r|\mathcal{F}_{t-1}) = \frac{r}{d(r, b_{t-1}, \theta)\sqrt{b_{t-1}^2 + 4r^2\varphi}}f_\epsilon(d(r, b_{t-1}, \theta))$$

with $d(r, b_{t-1}, \theta) = \sqrt{\frac{\sqrt{b_{t-1}^2+4r^2\varphi}-b_{t-1}}{2\varphi}}$. Note that $\epsilon_t = d(r_t; b_{t-1}, \theta_0)$. I now calculate the limit of the density function at $r = 0$. First observe that

$$\lim_{r\to 0}\frac{r}{d(r, b_{t-1}, \theta)} = \lim_{r\to 0}\frac{r}{\sqrt{\frac{(b_{t-1}^2+4\varphi r^2)^{\frac{1}{2}}-b_{t-1}}{2\varphi}}} = \lim_{r\to 0}\frac{r}{\sqrt{\left(\frac{b_{t-1}^2}{4\varphi^2} + \frac{4\varphi r^2}{4\varphi^2}\right)^{1/2} - \frac{b_{t-1}}{2\varphi}}} =$$

$$= \lim_{r\to 0}\frac{r}{\sqrt{\frac{b_{t-1}}{2\varphi}\left(\left(1 + \frac{4\varphi r^2}{b_{t-1}^2}\right)^{1/2} - 1\right)}} = \lim_{r\to 0}\frac{r}{\sqrt{\frac{b_{t-1}}{2\varphi}\left(1 + \frac{1}{2}\frac{4\varphi r^2}{b_{t-1}^2} - 1\right)}} = \sqrt{b_{t-1}}.$$

And as a result I have the following limit

$$\lim_{r\to 0} f_r(r|\mathcal{F}_{t-1}) = \lim_{r\to 0}\frac{r}{d(r, b_{t-1}, \theta)\sqrt{b_{t-1}^2 + 4r^2\varphi}}f_\epsilon(d(r, b_{t-1}, \theta)) = \frac{1}{\sqrt{b_{t-1}}}f_\epsilon(0).$$

The corresponding cumulative distribution function is given by

$$F(r|\mathcal{F}_{t-1}) = \int_{-\infty}^{d(r,b_{t-1},\theta)} f_\epsilon(x)dx = F_\epsilon\left(d(r, b_{t-1}, \theta)\right).$$

The j^{th} conditional moment of returns can be derived as follows (for the ease of exposition

I write $d(r)$ instead of $d(r, b_{t-1}, \theta)$):

$$E\left[r^j|\mathcal{F}_{t-1}\right] = \int_{-\infty}^{\infty} r^j f_r(r|\mathcal{F}_{t-1})\mathrm{d}r = \int_{-\infty}^{\infty} r^j \frac{r}{d(r)\sqrt{b_{t-1}^2 + 4r^2\varphi}} f_\epsilon(d(r))\mathrm{d}r =$$

$$= \int_{-\infty}^{\infty} r^j \frac{\sqrt{b_{t-1} + \varphi d(r)^2}}{\sqrt{b_{t-1}^2 + 4r^2\varphi}} f_\epsilon(d(r))\mathrm{d}r = \int_{-\infty}^{\infty} r^j \frac{\sqrt{b_{t-1} + \varphi d(r)^2}}{b_{t-1} + 2\varphi d(r)^2} f_\epsilon(d(r))\mathrm{d}r. \quad (1.26)$$

Now observe that

$$\mathrm{d}r = \mathrm{d}(d(r)) \frac{b_{t-1} + 2\varphi d(r)^2}{\sqrt{b_{t-1} + \varphi d(r)^2}}.$$

Thus with a change of variable of integration eq.(1.26) can be written as

$$E\left[r^j|\mathcal{F}_{t-1}\right] = \int_{-\infty}^{\infty} r^j \frac{\sqrt{b_{t-1} + \varphi d(r)^2}}{b_{t-1} + 2\varphi d(r)^2} f_\epsilon(d(r))\mathrm{d}r = \int_{-\infty}^{\infty} r^j f_\epsilon(d(r))\mathrm{d}(d(r)) =$$

$$= \int_{-\infty}^{\infty} d(r)^j \left(b_{t-1} + \varphi d(r)^2\right)^{j/2} f_\epsilon(d(r))\mathrm{d}(d(r)) = \int_{-\infty}^{\infty} d(r)^j \left(b_{t-1}(1 + \frac{\varphi d(r)^2}{b_{t-1}})\right)^{j/2} f_\epsilon(d(r))\mathrm{d}(d(r)) =$$

$$= \int_{-\infty}^{\infty} d(r)^j b_{t-1}^{j/2} \left((1 + \frac{\varphi d(r)^2}{b_{t-1}})\right)^{j/2} f_\epsilon(d(r))\mathrm{d}(d(r)) = b_{t-1}^{j/2} \times$$

$$\int_{-\infty}^{\infty} d(r)^j \left(1 + \frac{j}{2}\frac{\varphi d(r)^2}{b_{t-1}} + \frac{j}{2}\left(\frac{j}{2} - 1\right)\frac{1}{2!}\frac{\varphi^2 d(r)^4}{b_{t-1}^2} + \frac{j}{2}\left(\frac{j}{2} - 1\right)\left(\frac{j}{2} - 2\right)\frac{1}{3!}\frac{\varphi^3 d(r)^6}{b_{t-1}^3} + \cdots\right) f_\epsilon(d(r))\mathrm{d}(d(r)) =$$

$$= \begin{cases} b_{t-1}^{j/2} \sum\limits_{i=-1}^{\infty} \frac{1}{(i+1)!}\left(\frac{\varphi}{b_{t-1}}\right)^{i+1} \left(\prod\limits_{s=0}^{i}\left(\frac{j}{2} - s\right)\right) E\left[\epsilon_t^{j+2(i+1)}\right], & \text{if } j \text{ is odd,} \\ b_{t-1}^{j/2} \sum\limits_{i=-1}^{j/2-1} \frac{1}{(i+1)!}\left(\frac{\varphi}{b_{t-1}}\right)^{i+1} \left(\prod\limits_{s=0}^{i}\left(\frac{j}{2} - s\right)\right) E\left[\epsilon_t^{j+2(i+1)}\right], & \text{if } j \text{ is even,} \end{cases}$$

where I conventionally take $\prod\limits_{s=0}^{-1}\left(\frac{j}{2} - s\right) = 1$ for the product over an empty index set. ■

Proof of Theorem 2. The general model is given by:

$$r_t = \lambda_t \epsilon_t \quad (1.27)$$

$$\lambda_t^2 = \alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi\epsilon_t^2. \quad (1.28)$$

Since the error term ϵ_t is i.i.d, it is then obvious that the error process $(\epsilon_t)_{t\in\mathbb{Z}}$ is always strictly stationary and ergodic. Thus, $(r_t)_{t\in\mathbb{Z}}$ is a strictly stationary process if $(\lambda_t)_{t\in\mathbb{Z}}$ is strictly stationary. Therefore, the task of deriving the strict stationarity conditions for the whole process $(r_t, \lambda_t)_{t\in\mathbb{Z}}$ can be reduced to deriving strict stationarity conditions for $(\lambda_t^2)_{t\in\mathbb{Z}}$, given by eq.(1.28). In order to proceed one needs either to assume the trivial σ-algebra $\mathcal{F}_0$ (and a probability measure μ_0) for the starting value λ_0^2 or to assume that the system extends infinitely far into the past. I proceed by implementing the former approach, defining:

$$\mathbb{P}[\lambda_0^2 \in \Gamma] = \mu_0(\Gamma) \qquad \forall \Gamma \in \mathcal{B} \quad \text{and} \quad \mu_0\left((0,\infty)\right) = 1, \tag{1.29}$$

where $\mathcal{B}$ denotes the Borel sets on $[0,\infty)$. In order to find strict stationarity conditions of λ_t^2 I next rewrite eq.(1.28) in the form of the stochastic difference equation $Y_{t+1} = A_t Y_t + B_t$, where Y_t, A_t and B_t are given by:

$$A_t = \beta + \gamma\epsilon_t^2, \qquad B_t = \alpha + \varphi\epsilon_{t+1}^2 \qquad \text{and} \qquad Y_t = \lambda_t^2 \tag{1.30}$$

Since sequences $(A_t)_{t\in\mathbb{N}}$ and $(B_t)_{t\in\mathbb{N}}$ are measurable transformations of the strictly stationary and ergodic process $(\epsilon_t)_{t\in\mathbb{N}}$ I can make use of the Theorem 3.5.8 of Stout (1974) to claim that these sequences are strictly stationary and ergodic as well as the sequence $\Psi = (A_t, B_t)_{t\in\mathbb{N}}$. Recursively plugging things in, we get:

$$\begin{aligned} Y_{t+1} &= A_tY_t + B_t = A_tA_{t-1}Y_{t-1} + A_tB_{t-1} + B_t = A_tA_{t-1}A_{t-2}Y_{t-2} + A_tA_{t-1}B_{t-2} + A_tB_{t-1} + B_t = \\ &= \cdots = \left(\prod_{i=0}^{t} A_{t-i}\right) Y_0 + \sum_{i=0}^{t}\left(\prod_{j=0}^{i-1} A_{t-j}\right) B_{t-i}, \end{aligned} \tag{1.31}$$

with the usual convention that $\prod_{j=0}^{-1} A_{t-j} = 1$ for the product over an empty index set. I denote by Y an arbitrary $\mathbb{R}$-valued random variable, which is defined on the same probability space as Ψ. Note that Y and Ψ should not necessarily be independent. The solution $y_t(Y,\Psi)$ of eq.(1.31) is then given by:

$$y_t(Y,\Psi) = \left(\prod_{i=0}^{t-1} A_i\right) Y_0 + \sum_{i=0}^{t-1}\left(\prod_{j=t-i}^{t-1} A_j\right) B_{t-i-1}.$$

Since I have shown earlier that the sequence $\Psi = (A_t, B_t)$ is strictly stationary and ergodic, I can now apply Theorem 1 of Brandt (1986) to deduce that $y_t(\Psi) = \sum_{i=0}^{\infty}\left(\prod_{j=t-i}^{t-1} A_j\right) B_{t-i-1}$, $t \in \mathbb{N}$ is strictly stationary solution if and only if the following conditions are satisfied:

$$\mathbb{P}(A_0 = 0) > 0$$

or

$$-\infty \leqq E\,\log|A_0| < 0 \qquad\qquad E\left(\,\log|B_0|\right)^+ < \infty,$$

where $x^+ = \max(0,x)$ for $x \in \mathbb{R}$. Plugging in the expressions for A_0 and B_0, given by eq.(1.30) I get the following strict stationarity conditions:

$$-\infty \leqq E\,\log\left|\beta + \gamma\epsilon_0^2\right| < 0 \qquad\qquad E\left(\log\left|\alpha + \varphi\epsilon_0^2\right|\right)^+ < \infty,$$

in addition to requiring that $\beta > 0, \gamma > 0$ and $\varphi \neq 0$. ■

Proof of Theorem 3.

Substituting eq.(1.8) into eq.(1.7) one gets:

$$E\left[\lambda_t^2\right] = \alpha + \beta E\left[\lambda_{t-1}^2\right] + \gamma \left\{ E\left[\lambda_{t-1}^2\right] + \varphi\left(E[\epsilon_t^4] - 1\right) \right\} + \varphi = \alpha + \varphi + \gamma\varphi\left(E[\epsilon_t^4] - 1\right) + E\left[\lambda_{t-1}^2\right](\beta + \gamma).$$

Then under the condition $\beta + \gamma < 1$ the process is weakly stationary and it's first unconditional moment, denoted by $E\left[\lambda_1^2\right]$ is given by:

$$E\left[\lambda_1^2\right] = \frac{\alpha + \varphi + \gamma\varphi\left(E[\epsilon_t^4] - 1\right)}{1 - (\beta + \gamma)}.$$

Note that $\alpha + \varphi + \gamma\varphi\left(E[\epsilon_t^4] - 1\right) = \alpha + \varphi(1 - \gamma) + \gamma\varphi E\left[\epsilon_t^4\right] > 0$ since $\gamma < 1$ due to $\beta + \gamma < 1$ and $(\alpha, \beta, \gamma, \varphi) \geq 0$. ■

Proof of Theorem 4.

The proof follows directly Theorem 3 and the fact that $E[r_t^2] = E[\lambda_t^2] + \varphi(E[\epsilon_t^4] - 1)$. The last claim of Theorem 4 follows from Theorem 1 that $E[r_t|\mathcal{F}_{t-1}] = 0$ for all t, then $cov(r_t, r_s|\mathcal{F}_{t-1}) = E\left[r_t r_s|\mathcal{F}_{t-1}\right] = r_s E\left[r_t|\mathcal{F}_{t-1}\right]$ for all $s < t$. ■

Proof of Theorem 5.

$$\begin{aligned} E[r_t^4] = E[\lambda_t^4\epsilon_t^4] &= E\left[(\alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi\epsilon_t^2)(\alpha + \beta\lambda_{t-1}^2 + \gamma r_{t-1}^2 + \varphi\epsilon_t^2)\epsilon_t^4\right] = \\ &= \alpha^2 E[\epsilon_t^4] + 2\alpha\beta E[\lambda_{t-1}^2]E[\epsilon_t^4] + +2\alpha\gamma E[r_{t-1}^2]E[\epsilon_t^4] + 2\alpha\varphi E[\epsilon_t^6] + \beta^2 E[\lambda_{t-1}^4]E[\epsilon_t^4] + \\ &+ 2\beta\gamma E[\lambda_{t-1}^2]E[r_{t-1}^2]E[\epsilon_t^4] + 2\beta\varphi E[\lambda_{t-1}^2]E[\epsilon_t^6] + \gamma^2 E[r_{t-1}^4]E[\epsilon_t^4] + 2\varphi\gamma E[r_{t-1}^2]E[\epsilon_t^6] + \varphi^2 E[\epsilon_t^8]. \end{aligned} \tag{1.32}$$

It can be shown that $E\left[\lambda_{t-1}^4\right] = \left(E\left[\lambda_{t-1}^2\right]\right)^2 + \varphi^2\eta/(1-\beta^2)$, where $\eta = E\left[\epsilon_t^4\right] - 1$. Then defining $\chi := \beta^2/(1-\beta^2)$ and rearranging eq. (1.32) it follows that:

$$\begin{aligned} E\left[r_t^4\right] &= (\alpha^2 + 2\eta\alpha\gamma\varphi + \varphi^2\chi\eta)E[\epsilon_t^4] + (2\alpha\beta + 2\alpha\gamma + 2\eta\beta\gamma\varphi)E[\lambda_{t-1}^2]E[\epsilon_t^4] + (2\alpha\varphi + 2\eta\varphi^2\gamma)E[\epsilon_t^6] + \\ &+ (\beta^2 + 2\beta\gamma)\left(E[\lambda_{t-1}^2]\right)^2 E[\epsilon_t^4] + (2\beta\varphi + 2\gamma\varphi)E[\lambda_{t-1}^2]E[\epsilon_t^6] + \gamma^2 E[r_t^4]E[\epsilon_t^4] + \varphi^2 E[\epsilon_t^8]. \end{aligned}$$

If r_t is fourth-order stationary ($E[r_t^4] = E[r_{t-1}^4]$), then

$$E[r_1^4] = \Bigg[(\alpha^2+2\eta\alpha\gamma\varphi+\varphi^2\chi\eta)E[\epsilon_t^4]+\varphi^2E[\epsilon_t^8]+(2\alpha\beta+2\alpha\gamma+2\eta\beta\gamma\varphi)E[\lambda_{t-1}^2]E[\epsilon_t^4]+(2\alpha\varphi+2\eta\varphi^2\gamma)E[\epsilon_t^6]+ \\ +(\beta^2+2\beta\gamma)\left(E[\lambda_{t-1}^2]\right)^2E[\epsilon_t^4]+(2\beta\varphi+2\gamma\varphi)E[\lambda_{t-1}^2]E[\epsilon_t^6]\Bigg]\left[1-\gamma^2E[\epsilon_t^4]\right]^{-1}$$

Since $E[r_1^4]$ must be positive, γ^2 must also satisfy:

$$1-\gamma^2E[\epsilon_t^4] > 0 \quad \Leftrightarrow \quad \gamma^2 < \frac{1}{E[\epsilon_t^4]}. \qquad \blacksquare$$

Proof of Theorem 6.

I start by writing down the RT-GARCH model with leverage and feedback:

$$r_t = \lambda_t\varepsilon_t$$

$$\lambda_t^2 = \alpha+\beta\lambda_{t-1}^2+\gamma_1r_{t-1}^2\mathbb{1}_{(r_t>0)}+\gamma_2r_{t-1}^2\mathbb{1}_{(r_t\le0)}+\varphi_1\epsilon_t^2\mathbb{1}_{(\epsilon_t>0)}+\varphi_2\epsilon_t^2\mathbb{1}_{(\epsilon_t\le0)}. \tag{1.33}$$

Denoting by $\kappa := E\left[\varepsilon_t^4\right]$ and $\eta := \kappa-1$ and following the same steps as in the proof of Theorems 3 and 4 it follows that:

$$E[r_t^2] = \alpha+\beta E[\lambda_{t-1}^2]+\gamma_1E[r_{t-1}^2|r_t>0]+\gamma_2E[r_{t-1}^2|r_t\le0]+\varphi_1E\left[\epsilon_t^4|\epsilon_t{>}0\right]+\varphi_2E\left[\epsilon_t^4|\epsilon_t\le0\right] \tag{1.34}$$

and

$$E[\lambda_t^2] = \alpha+\beta E[\lambda_{t-1}^2]+\gamma_1E[r_{t-1}^2|r_t>0]+\gamma_2E[r_{t-1}^2|r_t\le0]+\varphi_1E\left[\epsilon_t^2|\epsilon_t{>}0\right]+\varphi_2E\left[\epsilon_t^2|\epsilon_t\le0\right]. \tag{1.35}$$

Combining eq.(1.34)-(1.35) then yields:

$$E[r_t^2] = E[\lambda_t^2]+(\varphi_1+\varphi_2)\Big(E\left[\varepsilon_t^4\right]-1\Big) \tag{1.36}$$

In addition it also follows that the unconditional first moment of λ_t^2 is

$$E[\lambda_1^2] = \frac{\alpha+(\varphi_1+\varphi_2)\left[\eta(\gamma_1+\gamma_2)+1\right]}{1-(\beta+\gamma_1+\gamma_2)}. \tag{1.37}$$

Using eq.(1.34) and (1.35) I can write:

$$E\left[\lambda_{t+1}^2|\mathcal{F}_t\right] = \alpha+(\varphi_1+\varphi_2)\left[\eta(\gamma_1+\gamma_2)+1\right]+(\beta+\gamma_1+\gamma_2)E\left[\lambda_t^2|\mathcal{F}_t\right]. \tag{1.38}$$

Now note that from eq.(1.37) it holds that:

$$\alpha + (\varphi_1 + \varphi_2)\left[\eta(\gamma_1 + \gamma_2) + 1\right] = \left[1 - (\beta + \gamma_1 + \gamma_2)\right] E\left[\lambda_1^2\right],$$

which, when substituted back into eq.(1.38), together with eq.(1.36) provides us with the formula in Theorem 6. ■

1.10 Appendix B.

This Appendix presents some additional empirical results. I start with the plots of the error density forecasts for $h = 1$ day.

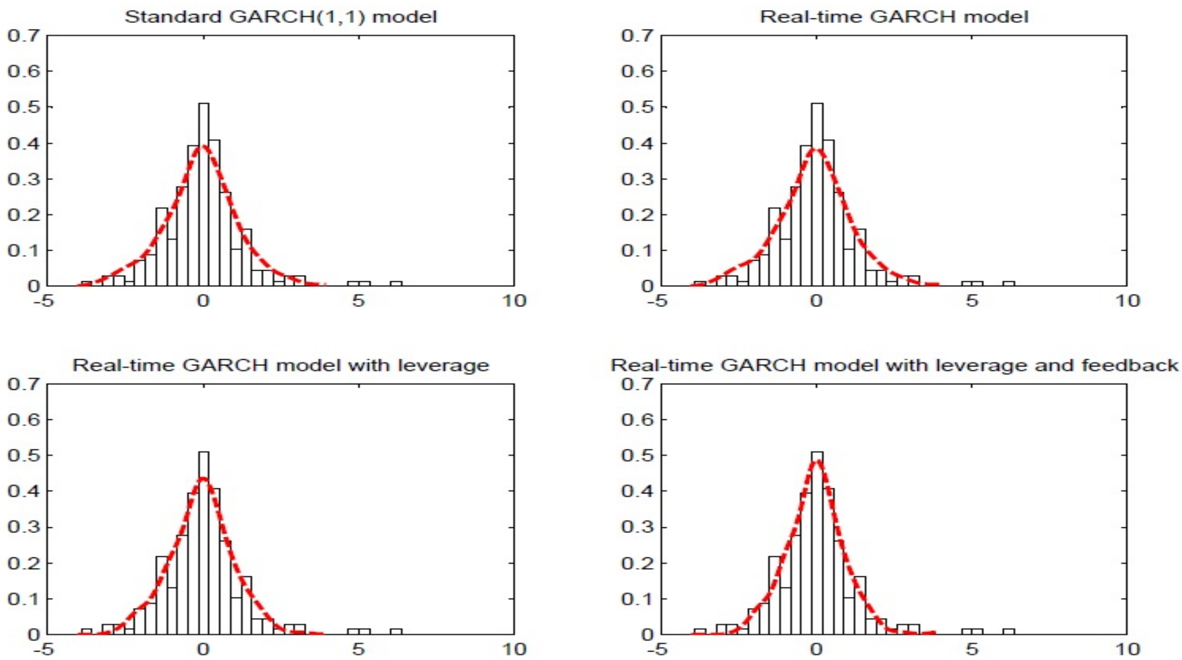


Figure 1.8: **Plot of the error density forecasts S&P 500** . The figure displays the error density forecasts. The histogram corresponds to $\epsilon_{t+l} = r_{t+l}/\sqrt{K(X_\sigma)_{t+l}}$, where $K(X_\sigma)$ is the realised kernel. The dashed line represents the kernel density estimator of the $\epsilon_{t+l} = r_{t+l}/\hat{\sigma}_{t+l}$ for the standard GARCH(1,1) model and $\epsilon_{t+l} = r_{t+l}/\hat{\lambda}_{t+l}$ for the Real-time GARCH models.

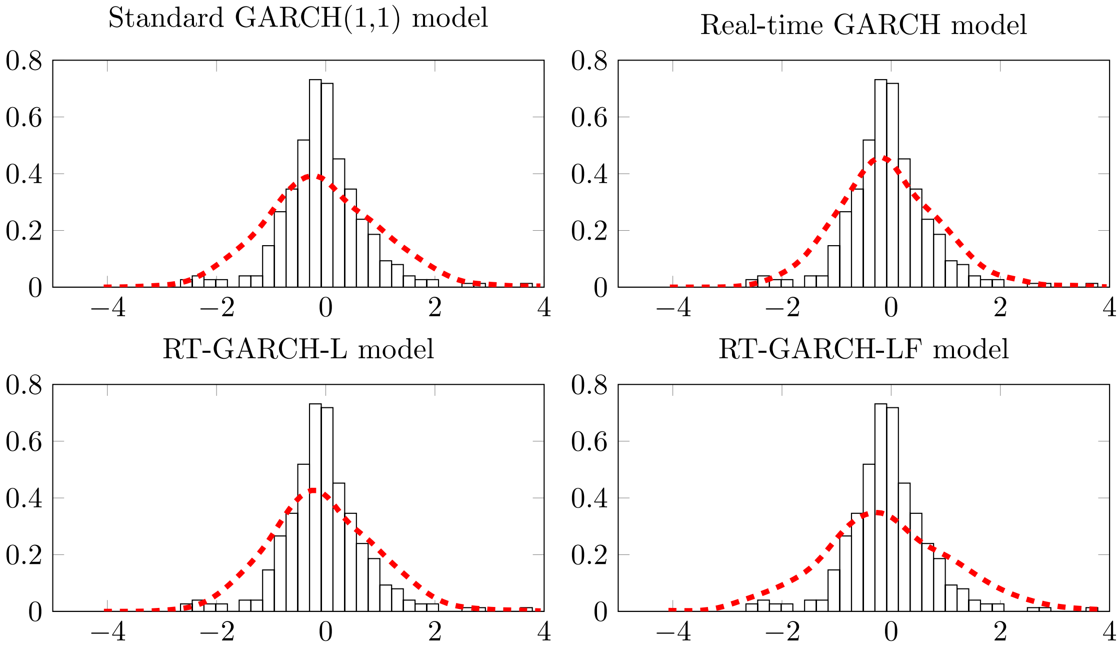


Figure 1.9: **Plot of the error density forecasts GE** . The figure displays the error density forecasts. The histogram corresponds to $\epsilon_{t+l} = r_{t+l}/\sqrt{K(X_\sigma)_{t+l}}$, where $K(X_\sigma)$ is the realised kernel. The dashed line represents the kernel density estimator of the $\epsilon_{t+l} = r_{t+l}/\hat{\sigma}_{t+l}$ for the standard GARCH(1,1) model and $\epsilon_{t+l} = r_{t+l}/\hat{\lambda}_{t+l}$ for the Real-time GARCH models.

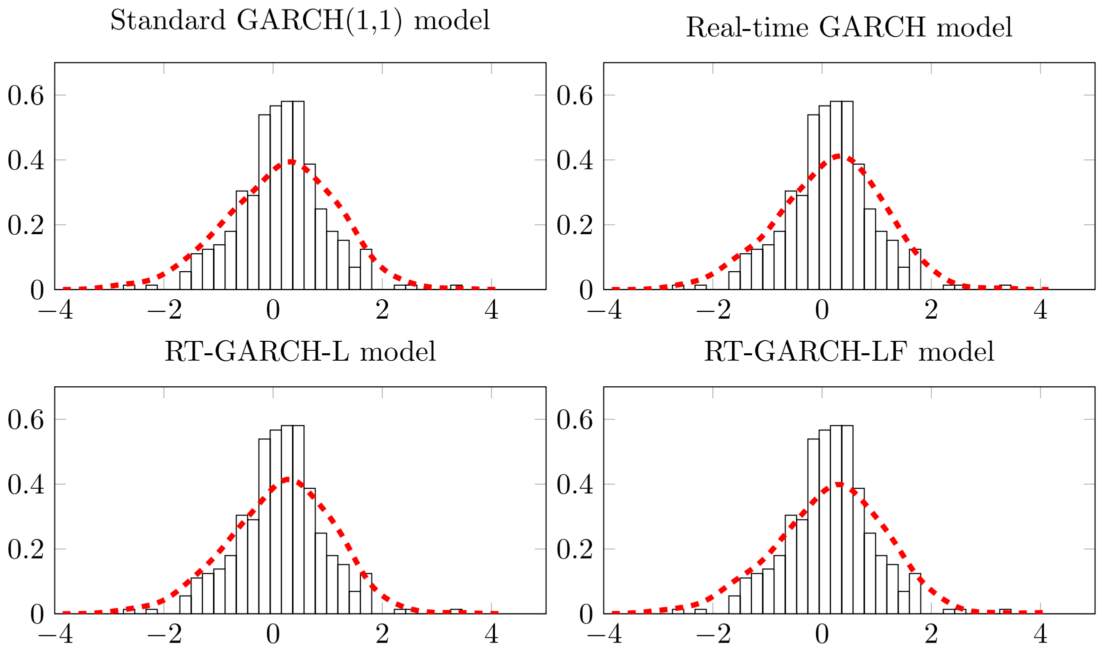


Figure 1.10: **Plot of the error density forecasts IBM** . The figure displays the error density forecasts. The histogram corresponds to $\epsilon_{t+l} = r_{t+l}/\sqrt{K(X_\sigma)_{t+l}}$, where $K(X_\sigma)$ is the realised kernel. The dashed line represents the kernel density estimator of the $\epsilon_{t+l} = r_{t+l}/\hat{\sigma}_{t+l}$ for the standard GARCH(1,1) model and $\epsilon_{t+l} = r_{t+l}/\hat{\lambda}_{t+l}$ for the Real-time GARCH models.

Table 1.16: Test for Superior Predictive ability (SPA) for IBM data .

Forecast horizon h		**h=1**			**h=5**			**h=10**			**h=15**	
Benchmark	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE
	SPA_l	0.4540	1	SPA_l	1	1	SPA_l	0.1070	0.0100	SPA_l	0.0040	0
Real-time GARCH	SPA_c	0.6000	1	SPA_c	1	1	SPA_c	0.1440	0.0140	SPA_c	0.0040	0
	SPA_u	0.6460	1	SPA_u	1	1	SPA_u	0.1540	0.0170	SPA_u	0.0040	0
	SPA_l	0.4320	0.5000	SPA_l	0.1000	0.3440	SPA_l	1	1	SPA_l	1	1
Real-time GARCH-L	SPA_c	0.4820	0.8400	SPA_c	0.1990	0.3440	SPA_c	1	1	SPA_c	1	1
	SPA_u	0.9110	0.9790	SPA_u	0.2510	0.8810	SPA_u	1	1	SPA_u	1	1
	SPA_l	0.0130	0.0020	SPA_l	0.0350	0.0110	SPA_l	0.0500	0.0100	SPA_l	0.0500	0.0080
Real-time GARCH-LF	SPA_c	0.0130	0.0200	SPA_c	0.0350	0.0110	SPA_c	0.0500	0.0100	SPA_c	0.0500	0.0080
	SPA_u	0.0150	0.0200	SPA_u	0.0370	0.0110	SPA_u	0.0540	0.0100	SPA_u	0.0580	0.0080
	SPA_l	1	0.0950	SPA_l	0.0780	0.0160	SPA_l	0.0150	0.0030	SPA_l	0.0140	0.0090
A-PARCH(2,2) $-St.t$ distr.	SPA_c	1	0.1140	SPA_c	0.0780	0.0160	SPA_c	0.0150	0.0030	SPA_c	0.0140	0.0090
	SPA_u	1	0.1700	SPA_u	0.0790	0.0160	SPA_u	0.0150	0.0030	SPA_u	0.0150	0.0090
	SPA_l	0.0160	0.0410	SPA_l	0.0630	0.0200	SPA_l	0.1360	0.0780	SPA_l	0.1180	0.2630
GARCH(1,1)-$N(0,1)$	SPA_c	0.0160	0.0410	SPA_c	0.0800	0.0300	SPA_c	0.3190	0.1180	SPA_c	0.2460	0.3760
	SPA_u	0.0160	0.0530	SPA_u	0.0840	0.0320	SPA_u	0.4210	0.3520	SPA_u	0.3040	0.5650
	SPA_l	0.0130	0.0580	SPA_l	0.0950	0.0210	SPA_l	0.1720	0.0780	SPA_l	0.1190	0.2310
GARCH(1,2)-$N(0,1)$	SPA_c	0.0130	0.0590	SPA_c	0.1200	0.0240	SPA_c	0.3580	0.1180	SPA_c	0.2710	0.3690
	SPA_u	0.0120	0.0680	SPA_u	0.1260	0.0260	SPA_u	0.4610	0.3520	SPA_u	0.3280	0.5650
	SPA_l	0.0450	0.0570	SPA_l	0.0450	0.0380	SPA_l	0.0960	0.0520	SPA_l	0.0940	0.3140
GARCH(1,1)-$St.t$ distr.	SPA_c	0.0460	0.0700	SPA_c	0.0480	0.0600	SPA_c	0.1040	0.0660	SPA_c	0.0940	0.3790
	SPA_u	0.0480	0.1070	SPA_u	0.0480	0.0670	SPA_u	0.1080	0.1120	SPA_u	0.1260	0.6350
	SPA_l	0.0340	0.0780	SPA_l	0.0560	0.0370	SPA_l	0.0710	0.0490	SPA_l	0.0750	0.2620
GARCH(1,2)-$St.t$ distr.	SPA_c	0.0340	0.0840	SPA_c	0.0610	0.0700	SPA_c	0.0990	0.0590	SPA_c	0.0910	0.4710
	SPA_u	0.0340	0.1010	SPA_u	0.0610	0.0900	SPA_u	0.1000	0.1200	SPA_u	0.1140	0.6480
	SPA_l	0	0	SPA_l	0.0010	0	SPA_l	0	0	SPA_l	0.0020	0.0020
Simple NoVaS	SPA_c	0	0	SPA_c	0.0010	0	SPA_c	0	0	SPA_c	0.0020	0.0020
	SPA_u	0	0	SPA_u	0.0010	0	SPA_u	0	0	SPA_u	0.0020	0.0020
	SPA_l	0	0	SPA_l	0.0020	0	SPA_l	0.0020	0	SPA_l	0.0020	0
Exponential NoVaS	SPA_c	0	0	SPA_c	0.0020	0	SPA_c	0.0020	0	SPA_c	0.0020	0
	SPA_u	0	0	SPA_u	0.0020	0	SPA_u	0.0020	0	SPA_u	0.0020	0

Note: Table reports the p-value of Hansen's (2005) Superior Predictive Ability test statistics (SPA_c) as well as its lower (SPA_l) and upper (SPA_u) bounds for different forecast horizons.

Table 22: Test for Superior Predictive ability (SPA) for GE data .

Forecast horizon h		**h=1**			**h=5**			**h=10**			**h=15**	
Benchmark	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE
	SPA_l	1	1	SPA_l	1	1	SPA_l	1	1	SPA_l	1	0.1540
Real-time GARCH	SPA_c	1	1	SPA_c	1	1	SPA_c	1	1	SPA_c	1	0.2440
	SPA_u	1	1	SPA_u	1	1	SPA_u	1	1	SPA_u	1	0.4020
	SPA_l	0.1870	0.0970	SPA_l	0.0450	0.0200	SPA_l	0.0540	0.1410	SPA_l	0.3140	1
Real-time GARCH-L	SPA_c	0.2740	0.1150	SPA_c	0.0450	0.0200	SPA_c	0.0590	0.1970	SPA_c	0.4720	1
	SPA_u	03030	0.2020	SPA_u	0.0640	0.0200	SPA_u	0.0880	0.4470	SPA_u	0.8910	1
	SPA_l	0.2730	0.2170	SPA_l	0.1210	0.1050	SPA_l	0.0350	0.0170	SPA_l	0.0420	0.0160
Real-time GARCH-LF	SPA_c	0.3790	0.3340	SPA_c	0.1510	0.1510	SPA_c	0.0430	0.0270	SPA_c	0.0550	0.0160
	SPA_u	0.5020	0.4980	SPA_u	0.2410	0.2440	SPA_u	0.0480	0.0410	SPA_u	0.0630	0.0350
	SPA_l	0.1890	0.0580	SPA_l	0.0140	0.0030	SPA_l	0.0160	0.0130	SPA_l	0.0120	0.0290
A-PARCH(2,2) $-St.t$ distr.	SPA_c	0.2450	0.0580	SPA_c	0.0140	0.0030	SPA_c	0.0160	0.0130	SPA_c	0.0140	0.0300
	SPA_u	0.3020	0.1130	SPA_u	0.0140	0.0040	SPA_u	0.0160	0.0140	SPA_u	0.0140	0.0380
	SPA_l	0.0260	0.0040	SPA_l	0.0210	0	SPA_l	0.0080	0.0030	SPA_l	0.0090	0.0060
GARCH(1,1)-$N(0,1)$	SPA_c	0.0260	0.0050	SPA_c	0.0210	0	SPA_c	0.0080	0.0030	SPA_c	0.0090	0.0060
	SPA_u	0.0320	0.0050	SPA_u	0.0210	0	SPA_u	0.0080	0.0030	SPA_u	0.0090	0.0060
	SPA_l	0.1140	0.0340	SPA_l	0.0210	0.0020	SPA_l	0.0100	0.0050	SPA_l	0.0230	0.0280
GARCH(1,2)-$N(0,1)$	SPA_c	0.1140	0.0390	SPA_c	0.0210	0.0020	SPA_c	0.0100	0.0050	SPA_c	0.0230	0.0320
	SPA_u	0.1700	0.0630	SPA_u	0.0240	0.0020	SPA_u	0.0100	0.0090	SPA_u	0.0250	0.0580
	SPA_l	0.0320	0.0070	SPA_l	0.0140	0	SPA_l	0.0090	0.0030	SPA_l	0.0100	0.0060
GARCH(1,1)-$St.t$ distr.	SPA_c	0.0320	0.0080	SPA_c	0.0140	0	SPA_c	0.0090	0.0030	SPA_c	0.0100	0.0060
	SPA_u	0.0350	0.0090	SPA_u	0.0150	0	SPA_u	0.0090	0.0030	SPA_u	0.0100	0.0060
	SPA_l	0.0310	0.0070	SPA_l	0.0130	0.0020	SPA_l	0.0160	0.0010	SPA_l	0.0190	0.0010
GARCH(1,2)-$St.t$ distr.	SPA_c	0.0310	0.0070	SPA_c	0.0130	0.0020	SPA_c	0.0160	0.0010	SPA_c	0.0190	0.0010
	SPA_u	0.0330	0.0080	SPA_u	0.0130	0.0020	SPA_u	0.0160	0.0010	SPA_u	0.0190	0.0010
	SPA_l	0.0020	0	SPA_l	0.0030	0	SPA_l	0.0030	0	SPA_l	0.0110	0
Simple NoVaS	SPA_c	0.0020	0	SPA_c	0.0030	0	SPA_c	0.0030	0	SPA_c	0.0110	0
	SPA_u	0.0020	0	SPA_u	0.0030	0	SPA_u	0.0030	0	SPA_u	0.0110	0
	SPA_l	0	0	SPA_l	0	0	SPA_l	0.0010	0	SPA_l	0.0020	0
Exponential NoVaS	SPA_c	0	0	SPA_c	0	0	SPA_c	0.0010	0	SPA_c	0.0020	0
	SPA_u	0	0	SPA_u	0	0	SPA_u	0.0010	0	SPA_u	0.0020	0

Note: Table reports the p-value of Hansen's (2005) Superior Predictive Ability test statistics (SPA_c) as well as its lower (SPA_l) and upper (SPA_u) bounds for different forecast horizons.

Table 23: Test for Superior Predictive ability (SPA) for S&P 500 data .

Forecast horizon h		**h=1**			**h=5**			**h=10**			**h=15**	
Benchmark	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE	p	MSE	QLIKE
Real-time GARCH	SPA_l	0.016	0.004	SPA_l	0.111	0.054	SPA_l	0.165	0.095	SPA_l	0.123	0.059
	SPA_c	0.024	0.004	SPA_c	0.176	0.054	SPA_c	0.316	0.103	SPA_c	0.267	0.08
	SPA_u	0.025	0.004	SPA_u	0.176	0.072	SPA_u	0.316	0.143	SPA_u	0.31	0.116
Real-time GARCH-L	SPA_l	0.102	0.019	SPA_l	0.144	0.101	SPA_l	0.177	0.246	SPA_l	0.434	1
	SPA_c	0.174	0.021	SPA_c	0.255	0.186	SPA_c	0.49	0.466	SPA_c	0.837	1
	SPA_u	0.174	0.028	SPA_u	0.255	0.278	SPA_u	0.496	0.604	SPA_u	0.922	1
Real-time GARCH-LF	SPA_l	1	1	SPA_l	0.309	0.302	SPA_l	0.066	0.281	SPA_l	0.023	0.256
	SPA_c	1	1	SPA_c	0.627	0.508	SPA_c	0.081	0.339	SPA_c	0.023	0.264
	SPA_u	1	1	SPA_u	0.628	0.605	SPA_u	0.081	0.381	SPA_u	0.023	0.323
A-PARCH(2,2) $-St.t$ distr.	SPA_l	0.367	0.277	SPA_l	1	1	SPA_l	1	1	SPA_l	1	0.395
	SPA_c	0.525	0.311	SPA_c	1	1	SPA_c	1	1	SPA_c	1	0.77
	SPA_u	0.667	0.584	SPA_u	1	1	SPA_u	1	1	SPA_u	1	0.894
GARCH(1,1)-$N(0,1)$	SPA_l	0.016	0.009	SPA_l	0.03	0.026	SPA_l	0.132	0.089	SPA_l	0.096	0.081
	SPA_c	0.018	0.009	SPA_c	0.054	0.026	SPA_c	0.172	0.103	SPA_c	0.194	0.125
	SPA_u	0.018	0.012	SPA_u	0.054	0.034	SPA_u	0.172	0.148	SPA_u	0.194	0.158
GARCH(1,2)-$N(0,1)$	SPA_l	0.024	0.003	SPA_l	0.141	0.04	SPA_l	0.181	0.066	SPA_l	0.089	0.038
	SPA_c	0.026	0.003	SPA_c	0.215	0.044	SPA_c	0.236	0.066	SPA_c	0.128	0.038
	SPA_u	0.029	0.004	SPA_u	0.215	0.062	SPA_u	0.236	0.082	SPA_u	0.128	0.052
GARCH(1,1)-$St.t$ distr.	SPA_l	0.018	0.013	SPA_l	0.077	0.017	SPA_l	0.135	0.023	SPA_l	0.088	0.027
	SPA_c	0.024	0.013	SPA_c	0.092	0.019	SPA_c	0.145	0.038	SPA_c	0.115	0.029
	SPA_u	0.024	0.018	SPA_u	0.092	0.022	SPA_u	0.145	0.04	SPA_u	0.115	0.034
GARCH(1,2)-$St.t$ distr.	SPA_l	0.027	0.005	SPA_l	0.091	0.013	SPA_l	0.117	0.026	SPA_l	0.057	0.034
	SPA_c	0.031	0.005	SPA_c	0.1	0.013	SPA_c	0.132	0.028	SPA_c	0.069	0.034
	SPA_u	0.031	0.006	SPA_u	0.1	0.014	SPA_u	0.132	0.03	SPA_u	0.069	0.035
Simple NoVaS	SPA_l	0	0	SPA_l	0	0	SPA_l	0	0	SPA_l	0	0
	SPA_c	0	0	SPA_c	0	0	SPA_c	0	0	SPA_c	0	0
	SPA_u	0	0	SPA_u	0	0	SPA_u	0	0	SPA_u	0	0
Exponential NoVaS	SPA_l	0	0	SPA_l	0	0	SPA_l	0	0	SPA_l	0	0
	SPA_c	0	0	SPA_c	0	0	SPA_c	0	0	SPA_c	0	0
	SPA_u	0	0	SPA_u	0	0	SPA_u	0	0	SPA_u	0	0

Note: Table reports the p-value of Hansen's (2005) Superior Predictive Ability test statistics (SPA_c) as well as its lower (SPA_l) and upper (SPA_u) bounds for different forecast horizons.

Chapter 2

Asymptotic Inference for Real-time GARCH(1,1) model

2.1 Introduction

Simplicity in formulation, estimation, and inference has greatly contributed to the standard GARCH-type models' popularity among practitioners. Yet it is well-known that GARCH models are poorly suited for situations of rapid changes in financial markets, see e.g. Andersen et al. (2003), and Hansen et al. (2011). This, in a large part, happens due to the fact that GARCH-type models consider volatility as a function of past information only, resulting in an inefficient use of the available information. While during calm periods this loss in information might not be too severe, it can drastically worsen volatility forecasts during times of rapid changes in financial markets. On a separate note, another implication of all standard GARCH models is that the standardised conditional moments of returns of order larger than $k > 2$ are constant. In particular, in GARCH models the conditional distribution of returns possesses a constant kurtosis. It is however reasonable to consider that the shape of the conditional distribution of returns changes with time, which is particularly relevant in times of turmoil.

In Chapter 1, a new model was developed, the Real-time GARCH (RT-GARCH thereafter), which addresses both of the above mentioned issues. Firstly, I show that it is possible to efficiently utilise all available information in GARCH models. In particular, the new model is given by[1]:

$$r_t = \varepsilon_t \sqrt{h_t}, \qquad \text{is} \quad \mathcal{F}_t - \text{measurable}, \tag{2.1}$$

where ε_t is assumed to be an independent and identically distributed (i.i.d.) error term satisfying $E[\varepsilon_t] = 0$ and $E\left[\varepsilon_t^2\right] = 1$, and $\mathcal{F}_{t-1} := \sigma(r_t, s \leq t-1)$ is the sigma-algebra induced by the history of returns up to time $t-1$. In eq.(2.1) h_t is a "volatility-like" process in the sense that it can be

[1] I slightly change the notation from Chapter 1, primarily for the ease of exposition of the results to follow.

easily related to $E\left[r_t^2|\mathcal{F}_{t-1}\right]$, but is not a volatility in the common sense, i.e. $E\left[r_t^2|\mathcal{F}_{t-1}\right] \neq h_t$ as h_t is *not* independent of ε_t any longer. Compared to standard GARCH models, the new model uses all available information up to time t instead of time $t-1$. Compared to SV models, the information set $\mathcal{F}_t$ contains only one source of randomness shared by the returns and volatility processes, which allows for a variety of the estimation methods, QML being perhaps the most common. Therefore, the RT-GARCH model can be thought of as a link between GARCH and SV models, as it nests most GARCH-type models as its special case and can be interpreted as a special case of SV model, see Chapter 1 for a detailed discussion.

In addition, this new model allows the shape of the conditional distribution of returns to be time-varying. For example, the conditional density of returns is no longer a scaled normal density even when the error term has a Gaussian density. More precisely, the conditional density has an extra shape parameter characterising the "peakedness" and/or thickness of the tails of the returns' density. This allows the new model to be better capture the tail behaviour of the returns. As a result, more precise Value-at-Risk (VaR) forecasts as well as short- and long-run volatility forecasts can be obtained.

In Chapter 1 several theoretical properties of the model have been worked out, including weak and strict stationarity conditions, conditional density and distribution of returns implied by the model. As the conditional density of returns can be written in an analytical form, the estimation of the new model can be done via the QML. However, the results for the asymptotic inference for the QMLE are not provided. This chapter establishes these results. Since RT-GARCH(1,1) nests the standard GARCH(1,1) model, one might expect that the asymptotic theory for this model must be a generalisation of some sort of the existing asymptotic theory for the GARCH(1,1) model. Although this turns out to be true, the generalisation is not straightforward due to the added nonlinearity in the model. For instance, for GARCH models under the correct specification of the first two moments of returns, the score function is automatically a martingale difference sequence. For the RT-GARCH(1,1) model, establishing that the score function at the true parameter vector is a martingale difference sequence requires establishing some intermediate results. Similarly, due to the high nonlinearity of the score function, proving the finiteness of various moments, such as those of the log-likelihood, the score, and the higher order derivatives, poses certain challenges as well. In many ways, the theory presented here can be seen as a generalisation of the QMLE for the standard GARCH(1,1) model, developed, among others, by Francq and Zakoïan (2004). Conceptually the theory I present here is also similar to that developed by Kristensen and Han (2014) for the QMLE of the GARCH-X models. However, in my case, an extra variable in the volatility equation is endogenous, unobserved and $\mathcal{F}_t$-measurable, which complicates the analysis.

I start by establishing the ergodicity of the joint process $(r_t^2,)$. I then show that the score function is a martingale difference sequence that nests the GARCH(1,1) score as its special case. This is a nontrivial result as, in comparison with the standard GARCH(1,1) model, conditional on $\mathcal{F}_{t-1}$ the score function is not separable in h_t and ε_t any longer. Therefore the fact that the score function is a martingale difference function does not follow by construction but requires some additional proofs. I next prove the consistency of the parameter vector and asymptotic normality of the QMLE at the usual $\sqrt{T}$ rate. The generalisation of the theory requires higher number of moments for the error terms to exist. More precisely, I will require the assumption $E\left[|\varepsilon_t|^{10}\right] < \infty$ to hold, which is a stronger condition than for the standard GARCH(1,1) model where $k \leq 2$.

The remainder of the chapter is structured as follows. Section 2.2 presents the main results, which include strong consistency and asymptotic normality of the parameter vector, as well as a discussion of the ways these results are related to those of the standard GARCH(1,1) QMLE theory. Section 2.3 discusses the simulations and section 2.4 concludes. All proofs of Lemmas and Theorems in Section 2.2 can be found in the Appendix A and all simulation results are presented in the Appendix B. Throughout the paper $\xrightarrow{p}$ denotes the convergence in probability, $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and $||\cdot||_p$ for $p \geq 1$ denotes the L^p-norm on $(\Omega, \mathcal{F}, \mathbb{P})$.

2.2 Main Results

I start by introducing the necessary notation. Denote the parameter vector by $\theta := (w, \alpha, \beta, \varphi)' \in \Theta \subseteq \mathbb{R}^4$, where Θ is compact. For reasons that will become apparent later on it is useful to isolate the parameter φ from the whole vector. I therefore denote by $\lambda := (w, \alpha, \beta)'$, which then implies that $\theta = (\lambda', \varphi)'$. The true parameter vector is denoted by $\theta_0 = (\lambda_0', \varphi_0)'$. Moreover, for the rest of the chapter I make use of the following notation: I denote by $h_t(\theta)$ the general version of the volatility-like process, by $h_t := h_t(\theta_0)$ the true volatility-like process. Moreover, variables with $\star$ superscript will denote the stationary versions, i.e. $h_t^\star(\theta)$ denotes the stationary version of $h_t(\theta)$ process and $h_t^\star = h_t^\star(\theta_0)$ denotes the stationary version evaluated at the true parameter vector θ_0. Similar notation will be used for the log-likelihood and its derivatives. With this notation in hand the general model, introduced in Chapter 1, is then given by the following two equations:

$$r_t = \sqrt{h_t(\theta)}\varepsilon_t, \tag{2.2}$$

and

$$h_t(\theta) = \underbrace{w + \alpha r_{t-1}^2 + \beta h_{t-1}(\theta)}_{b_{t-1}(\theta)} + \varphi \varepsilon_t^2(\theta), \quad \text{where} \quad \varepsilon_t^2(\theta) = r_t^2 / h_t(\theta), \tag{2.3}$$

where $t \in \mathbb{Z}$ and ε_t is an i.i.d.(0,1) sequence. The true data-generating process is given by the following two equations:

$$r_t = \sqrt{h_t(\theta_0)}\varepsilon_t, \tag{2.4}$$

and

$$h_t = h_t(\theta_0) = \underbrace{w_0 + \alpha_0 r_{t-1}^2 + \beta_0 h_{t-1}(\theta_0)}_{b_{t-1}(\theta_0)} + \varphi_0 \varepsilon_t^2, \tag{2.5}$$

where I used that fact that $\varepsilon_t(\theta_0) = \varepsilon_t$ recovers the true error term. In addition, for a general parameter vector θ and the corresponding $b_{t-1}(\theta)$ the following holds (see Theorem 1 in Chapter 1):

$$\varepsilon(r_t, b_{t-1}(\theta)) = \begin{cases} sign(r_t)\sqrt{\frac{\sqrt{b_{t-1}^2(\theta)+4r_t^2\varphi}-b_{t-1}(\theta)}{2\varphi}}, & for \quad \varphi \neq 0 \\ r_t/\sqrt{b_{t-1}(\theta)}, & for \quad \varphi = 0 \end{cases} \tag{2.6}$$

where again setting $\theta = \theta_0$ implies $\varepsilon(r_t, b_{t-1}(\theta_0)) = \varepsilon_t$, recovering therefore the true error term in (2.2)-(2.3). The notation $\varepsilon(r_t, b_{t-1}(\theta))$ is used to indicate that ε_t can be expressed as a function of r_t, b_{t-1} and the parameter vector θ. However, for notational convenience and for the ease of exposition, for the rest of the chapter I write $\varepsilon_t(\theta)$ instead of $\varepsilon(r_t, b_{t-1}(\theta))$. I denote by $\mathcal{F}_t$ the natural filtration. The conditional density for returns, derived in Theorem 1 in Chapter 1, is given by:

$$f_r(r|\mathcal{F}_{t-1}) = \frac{r}{\varepsilon(\theta)\sqrt{b_{t-1}^2(\theta) + 4r^2\varphi}} f_\varepsilon(\varepsilon(\theta)) = \frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi\frac{r_t^2}{h_t(\theta)}} f_\varepsilon\left(\frac{r_t}{\sqrt{h_t(\theta)}}\right), \tag{2.7}$$

where $b_{t-1}(\theta) \equiv w + \alpha r_{t-1}^2 + \beta h_{t-1}(\theta)$ and $f_\varepsilon(\cdot)$ is the probability density function of ε_t. In the special case of $\varphi = 0$ eq.(2.7) reduces to the standard GARCH(1,1) model conditional density which is just re-scaled Gaussian density.

Furthermore, let r_t for $t = 1, 2, \cdots T$ be observations from eq.(2.2)-(2.3). I then consider the estimation of the parameter of interest θ using the Gaussian log-likelihood with $\varepsilon_t \sim i.i.d.(0,1)$. In this case the negative log-likelihood function is given by:

$$L_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} l_t(\theta), \tag{2.8}$$

where the negative log-likelihood at time t is given by:

$$l_t(\theta) = \frac{1}{2}\log(2\pi) + \frac{1}{2}\frac{r_t^2}{h_t(\theta)} - \log\left(\frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi\frac{r_t^2}{h_t(\theta)}}\right). \tag{2.9}$$

In addition, the process $h_t(\cdot)$ is assumed to be initialised at some fixed parameter-independent value $\bar{h}_0 > 0$ and $h_0(\theta) = \bar{h}_0$. Similarly, let $\varepsilon_1(\theta) = \bar{\varepsilon}_1$. Note that (2.9) is the true log-likelihood function if ε_t were indeed Gaussian. Throughout, I do not however restrict ε_t to be Gaussian, and therefore $L_T(\theta)$ should be interpreted as quasi-log-likelihood. The QMLE of θ is then defined as:

$$\widehat{\theta}_T = \arg\max_{\theta\in\Theta} \left(-L_T(\theta)\right) = \arg\min_{\theta\in\Theta} L_T(\theta).$$

I start by showing that at θ_0 there exists a stationary and ergodic solution to eq. (2.2)-(2.3). To do this I first re-write eq.(2.4)-(2.5) as a stochastic recurrence equation (SDE) of the form:

$$X_t = A_t X_{t-1} + B_t, \tag{2.10}$$

$$X_t = (r_t^2, h_t)' \qquad B_t = \begin{pmatrix} w_0\varepsilon_t^2 + \varphi_0\varepsilon_t^4 \\ w_0 + \varphi_0\varepsilon_t^2 \end{pmatrix} \qquad \text{and} \qquad A_t = \begin{pmatrix} \alpha_0\varepsilon_t^2 & \beta_0\varepsilon_t^2 \\ \alpha_0 & \beta_0 \end{pmatrix},$$

where again I write $h_t = h_t(\theta_0)$. Before stating the first result I introduce a necessary assumption, which will insure that the top Lyapunov exponent associated with the sequence $\{A_t,\ t \in \mathbb{Z}\}$, denoted by $\gamma(A_0)$, is strictly negative. Let $\|\cdot\|$ denote a norm on $\mathbb{R}^2$ and define the operator norm on the space of $\mathbb{R}^{2\times 2}$ by $\|M\| = \sup\{\|Mx\|/\|x\|;\ x \in \mathbb{R}^2,\ x \neq 0\}$, for any $M \in \mathbb{R}^{2\times 2}$.

Assumption 1.

(i) *The innovations ε_t are i.i.d.(0,1).*

(ii) $\mathbb{P}\left(A_0 = 0\right) = 0, \quad -\infty \leq E\left[\log\|A_0\|\right] < 0, \quad E\left[\log^+\|B_0\|\right] < \infty.$

(iii) $\kappa = E\left[\varepsilon_t^4\right] < \infty.$

Theorem 1. (Stationarity and ergodicity). *Let the joint process $X_t = (r_t^2, h_t(\theta))$ be defined by eq.(2.2)-(2.3). Under Assumption 1 there exists an a.s.-unique causal ergodic strictly stationary solution to eq.(2.2)-(2.3) at θ_0.*

Remark 1. Assumption 1 provides sufficient conditions for the top Lyapunov exponent associated with the sequence $\{A_t,\ t \in \mathbb{Z}\}$ to be strictly negative, i.e. $\gamma(A_0) < 0$, where

$$\gamma(A_0) := \lim_{t\to\infty} \frac{1}{t} \log\|A_0 A_{-1}\cdots A_{-t}\| \qquad a.s.$$

This, together with the rest of the Assumption 1, will guarantee the existence of an ergodic strictly stationary solution to (2.10), see Theorem 2.1.3 in Buraczewski et. al. (2016). The unique strictly stationary solution is then given by $X_t = \sum_{i=0}^{\infty}\left(\prod_{j=t-i}^{t-1} A_j\right) B_{t-i-1}$. Moreover, note that Assumption 1 implies that $\alpha_0 + \beta_0 < 1$ and therefore implies the existence of the first moment of $h_t(\theta)$ and existence of the second moment r_t.

In what follows we need to establish that the quantities $h_t(\theta), L_T(\theta), l_t(\theta)$ and $\partial l_t(\theta)/\partial\theta$ as well as $\partial^2 l_t(\theta)/\partial\theta\partial\theta'$ can be approximated by their respective stationary versions $h_t^\star(\theta), L_T^\star(\theta), l_t^\star(\theta)$ and $\partial l_t^\star(\theta)/\partial\theta$ as well as $\partial^2 l_t^\star(\theta)/\partial\theta\partial\theta'$. Note that unlike the standard GARCH(1,1) case eq. (2.3) for $h_t(\theta)$ involves $\varepsilon_t^2(\theta)$, and therefore

$$h_t(\theta) = w+\alpha r_{t-1}^2+\beta h_{t-1}(\theta)+\varphi\varepsilon_t^2(\theta) = w+\varphi\varepsilon_t^2(\theta)+\alpha r_{t-1}^2+\beta\left[w+\alpha r_{t-2}^2+\beta h_{t-2}(\theta)+\varphi\varepsilon_{t-1}^2(\theta)\right] =$$
$$\cdots = w\sum_{j=1}^{t}\beta^j+\alpha\sum_{j=0}^{t-1}\beta^j r_{t-1-j}^2+\varphi\sum_{j=0}^{t}\beta^j\varepsilon_{t-j}^2(\theta)+\beta^t h_0(\theta),$$

and given that for any j, $\varepsilon_{t-j}^2(\theta) = r_{t-j}^2/h_{t-j}(\theta)$, this results in highly nonlinear structure. To proceed further I therefore will make use of the following high-level assumption.

Assumption 2.

For all t it holds that $|\varepsilon_t^2(\theta)-\varepsilon_t^{2,\star}(\theta)| \le K\rho^t$ for some $\rho<1$.

Before I start the next main result, I need to introduce some additional necessary assumptions for developing the asymptotic theory. Some of them will be quite familiar to the reader from the existing asymptotic theory for GARCH(1,1) model, e.g. Lee and Hansen (1994), Lumsdaine (1996), Francq and Zakoïan (2004), Kristensen and Rahbek (2009) among others. However I show how some of the assumptions shall be modified to reflect my case. For example, I will require stronger conditions on the moments of the error terms. I discuss each assumption below.

Assumption 3.

(i) $\Theta = \{\theta : 0 < \underline{w} \le w \le \overline{w},\ 0 \le \overline{\alpha} \le \alpha \le \overline{\alpha},\ 0 \le \underline{\beta} \le \beta \le \overline{\beta},\ 0 \le \underline{\varphi} \le \varphi \le \overline{\varphi}\}$, *where* $\overline{w} < \infty$, $\overline{\alpha} < \infty$, $\overline{\beta} < 1$ *and* $\overline{\varphi} < \infty$. *The true parameter vector* $\theta_0 \in \Theta$, *and* $(\alpha_0, \varphi_0) \neq (0,0)$.

(ii) ε_t^2 *has a non-degenerate distribution.*

I discuss the above Assumptions in turn. Assumption 3(i) defines Θ to be a compact set. It also imposes that all the parameters are positive since this will be sufficient to ensure that the process h_t is positive. Note, however, that the condition $w>0$ is not strictly necessary as unlike the standard GARCH(1,1) case, $E[h_1] = (w+\varphi+\alpha(\kappa-1))/(1-\beta-\alpha)$ with $\kappa := E(\varepsilon_t^4)$ (see Theorem 3 in Chapter 1). It is therefore obvious that the requirement $w>0$ will rule out $E[h_1]=0$ even if the rest of the parameters are zero. However, unlike in the standard GARCH(1,1) model, one can instead impose the condition $\varphi>0$ allowing then $w \ge 0$. The requirement of

$(\alpha_0, \varphi_0) \neq (0,0)$ is necessary to ensure identification of β_0 since whenever $(\alpha_0, \varphi_0) = (0,0)$ then $(\theta_0) = w_0/(1-\beta_0)$ and therefore w_0 and β_0 are not separately identified. Assumption 3(ii) is also required for identification and makes sure that ε_t is not concentrated at ± 1. Note that in Chapter 1 I used the stronger assumption that ε_t has a density f_ε everywhere. For the purpose of the proofs in this chapter, it is possible to establish the results using a weaker assumption, hence the weaker assumption made for the sake of generality.

Lemma 1. *Let $L_T(\theta)$ be defined in eq.(2.8). Denote further by $L_T^\star$ the stationary sequence which is asymptotically equivalent to $L_T(\theta)$. Then under Assumptions 1-2 it holds that*

$$\sup_{\theta \in \Theta} \frac{1}{T} \left| L_T(\theta) - L_T^\star(\theta) \right| \xrightarrow{p} 0, \tag{2.11}$$

where

$$L_T^\star(\theta) = \frac{1}{T} \sum_{t=1}^{T} l_t^\star(\theta),$$

and

$$l_t^\star(\theta) = \frac{1}{2} \log(2\pi) + \frac{1}{2} \frac{r_t^2}{h_t^\star(\theta)} - \log \left(\frac{\sqrt{h_t^\star(\theta)}}{h_t^\star(\theta) + \varphi \frac{r_t^2}{h_t^\star(\theta)}} \right),$$

where $h_t^\star(\theta)$ denotes the stationary sequence that is asymptotically equivalent to $h_T(\theta)$. Note further that Assumption 1 (ii) ensures that the second moment of r_t exists and therefore $E[|l_t^\star(\theta)|] < \infty$. One can now appeal to the uniform Law of Large Numbers (LLN) for stationary and ergodic sequences to establish that $L_T(\theta) \xrightarrow{p} L^\star := E\left[l_t^\star(\theta)\right]$ uniformly in θ. In what follows, the rest of the chapter is presented under the assumption that we have observed the stationary solution. I am now in a position to state the next main result.

Theorem 2. (Strong consistency of the QMLE). Under Assumptions 1-3, the QMLE $\widehat{\theta}_T$ is consistent, i.e. almost surely,

$$\widehat{\theta}_T \to \theta_0, \qquad as \qquad T \to \infty.$$

I now turn to investigating the properties of the score function. Since it is not possible to write down the score as a function of the whole parameter vector θ, it is necessary to split the parameter vector into two subsets: $\lambda := (w, \alpha, \beta)$ and φ such that $\theta = (\lambda', \varphi)'$. I will write down the score function for each of the subsets and establish that it is a martingale difference sequence. In addition, for convenience of exposition I write $\frac{\partial L_T(\theta_0)}{\partial \theta}$ and $\frac{\partial l_t(\theta_0)}{\partial \theta}$ to denote the derivative of $L_T(\theta)$ and $l_t(\theta)$ with respect to θ evaluated at the true parameter vector θ_0. In addition for the ease of the exposition of the theorems to follow I introduce the several useful quantities in the Lemma 2 below.

Lemma 2. *Let $\dot{h}_{t,\lambda}(\theta) := \partial h_t(\theta)/\partial\lambda$ and $\dot{h}_{t,\varphi}(\theta) := \partial h_t(\theta)/\partial\varphi$ and similarly $\dot{b}_{t-1,\lambda}(\theta) := \partial b_{t-1}(\theta)/\partial\lambda$ and $\dot{b}_{t-1,\varphi}(\theta) := \partial b_{t-1}(\theta)/\partial\varphi$. The following holds:*

$$\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)} = \frac{\varepsilon_t^2(\theta)}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)},$$

and

$$\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} = \frac{\dot{b}_{t-1,\lambda}(\theta)}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)},$$

where the elements of $\dot{b}_{t-1,\lambda}(\theta) = \left(\dot{b}_{t-1,w}(\theta), \dot{b}_{t-1,\alpha}(\theta), \dot{b}_{t-1,\beta}(\theta)\right)'$ are given by:

$$\dot{b}_{t-1,w}(\theta) = 1 + \sum_{k=0}^{t-1}\beta^{k+1}\prod_{j=0}^{k}\frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)},$$

$$\dot{b}_{t-1,\alpha}(\theta) = \sum_{k=0}^{t-1}\beta^k r_{t-1-k}^2\prod_{j=0}^{k-1}\frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)},$$

and

$$\dot{b}_{t-1,\beta}(\theta) = \sum_{k=0}^{t-1}\beta^k h_{t-1-k}(\theta)\prod_{j=0}^{k-1}\frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)}.$$

It also holds that

$$\dot{b}_{t-1,\varphi}(\theta) = \sum_{k=1}^{t-1}\beta^k\prod_{j=1}^{k}\frac{\varepsilon_{t-j}^2(\theta)h_{t-j}(\theta)}{h_{t-j}(\theta) + \varphi\varepsilon_{t-j}^2(\theta)}.$$

For establishing that the partial derivatives can be well approximated by their stationary versions I will assume the following high-level assumptions.

Assumption 4.
Let the negative log-likelihood function at time t, $l_t(\theta)$, be defined by eq. (2.9). For all t it holds:

(i) $\left\|\frac{\partial l_t(\theta)}{\partial\lambda} - \frac{\partial l_t^\star(\theta)}{\partial\lambda}\right\| \leq K_1\rho_1^t$ *for some* $\rho_1 < 1$,

(ii) $\left|\frac{\partial l_t(\theta)}{\partial\varphi} - \frac{\partial l_t^\star(\theta)}{\partial\varphi}\right| \leq K_2\rho_2^t$ *for some* $\rho_2 < 1$,

where K_1 and K_2 denote generic constants.

We are now in the position to state the next main result: the score function evaluated at the true parameter vector θ_0 is a martingale difference sequence.

Theorem 3. *Let ε_t be i.i.d.(0,1) random variables and let $(r_t,)$ evolve according to eq.(2.2)-(2.3), and assume Assumptions 1-4 hold. Let $\mathcal{F}_{t-1} = \sigma(r_t, s \leq t-1)$ be the σ-algebra induced by the history of returns up to time $t-1$. Denote by $\lambda = (w, \alpha, \beta)'$ and $\theta := (\lambda', \varphi)'$, in addition denote the true parameter vector by $\theta_0 = (\lambda_0', \varphi_0)'$ $\lambda_0 = (w_0, \alpha_0, \beta_0)$, and write $\varepsilon_t(\theta) = r_t^2/h_t(\theta)$. The score function is given by:*

$$\frac{\partial L_T(\theta)}{\partial \theta} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial l_t(\theta)}{\partial \theta},$$

where

$$\frac{\partial l_t(\theta)}{\partial \lambda} = \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)[h_t(\theta) + \varphi\varepsilon_t^2(\theta)]}\Bigg(-\frac{1}{2}h_t(\theta)[1-\varepsilon_t^2(\theta)] + \frac{1}{2}\varphi\left[3\varepsilon_t^2(\theta) + \varepsilon_t^4(\theta)\right]\Bigg),$$

and

$$\begin{aligned}\frac{\partial l_t(\theta)}{\partial \varphi} &= \frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)\left[h_t(\theta) + \varphi\varepsilon_t^2(\theta)\right]}\Bigg(-\frac{1}{2}h_t(\theta)[3-\varepsilon_t^2(\theta)] + \frac{1}{2}\varphi[\varepsilon_t^4(\theta) + \varepsilon_t^2(\theta)]\Bigg) = \\ &= \frac{\varepsilon_t^2(\theta)}{\left(h_t(\theta) + \varphi\varepsilon_t^2(\theta)\right)^2}\Bigg(-\frac{1}{2}h_t(\theta)[3-\varepsilon_t^2(\theta)] + \frac{1}{2}\varphi[\varepsilon_t^4(\theta) + \varepsilon_t^2(\theta)]\Bigg),\end{aligned}$$

where $\dot{h}_{t,\varphi}(\theta)/h_t(\theta)$ and $\dot{h}_{t,\lambda}(\theta)/h_t(\theta)$ are defined in Lemma 2. It then holds that:

$$E\left[\frac{\partial l_t^\star(\theta_0)}{\partial \lambda}\middle|\mathcal{F}_{t-1}\right] = 0 \qquad \text{and} \qquad E\left[\frac{\partial l_t^\star(\theta_0)}{\partial \varphi}\middle|\mathcal{F}_{t-1}\right] = 0.$$

The proof can be found in the Appendix A.

Remark 2. Note that this is a generalisation of the standard GARCH(1,1) result, which is a nested model. In particular, recall that the standard GARCH(1,1) model is obtained by setting $\varphi = 0$ resulting in the following expression:

$$\frac{\partial L_T(\theta)}{\partial \theta} = -\frac{1}{T}\sum_{t=1}^{T}\frac{1}{b_{t-1}(\theta)}\frac{\partial b_{t-1}(\theta)}{\partial \theta}(1-\varepsilon_t^2(\theta)), \qquad \text{where} \qquad \varepsilon_t^2(\theta) = r_t^2/b_{t-1}(\theta). \tag{2.12}$$

which is the GARCH(1,1) score function. Taking expectation of (2.12) makes it immediately obvious that the score function is a martingale difference function. However, establishing the same result for the score function of the RT-GARCH(1,1) model requires establishing some intermediate results, see Lemmas 4-5 in Appendix A. Once the fact that the score function is a martingale difference sequence is established, one can verify the finiteness of the moments of the score function and apply the standard martingale theory, e.g. Hall and Heyde (1980) or Pollard (1984) to establish asymptotic normality. I next turn to establishing the asymptotic normality of the QMLE $\widehat{\theta}$. Before doing so, however, some further necessary assumptions are needed.

Assumption 5.

(i) $\theta_0 \in \mathring{\Theta}$, *where* $\mathring{\Theta}$ *denotes the interior of* Θ.

(ii) $E\left[|\varepsilon_t|^{10}\right] < \infty$.

Assumption 5(i) ensures that the true parameter vector θ_0 is not on the boundary of the parameter space Θ. In particular, if any of the elements of the parameter vector θ_0, say $\theta_{0i} = 0$, then $\sqrt{T}\left(\widehat{\theta}_{0i} - \theta_{0i}\right) = \sqrt{T}\widehat{\theta}_{0i} \geq 0$ for all T and therefore will have a non-Gaussian distribution. I believe that the distribution of the parameters on the boundary will be similar to that in the standard GARCH case, developed in Francq and Zakoïan (2007). In this paper, however, I focus on developing the asymptotic theory for the parameter vector in the interior of Θ and leave the boundary case for future research. Finally Assumption 5(ii) is a moment condition on the error term, and here is where I need a stronger condition than for the standard GARCH(1,1) model, where it is required that $E\left[|\varepsilon_t|^4\right] < \infty$. The necessity of this assumption also comes from the fact that the score vector and the hessian matrix are non-singular and since eq.(2.3) invokes an additional term ε_t^2 (producing ε_t^4 in the r_t^2 equation), it becomes apparent that this will require more moments than in the standard GARCH(1,1) case. Before stating the last main result, I need to introduce yet another quantity: $\ddot{h}_{t,\lambda,\lambda}(\theta) := \partial^2 h_t(\theta)/\partial\lambda\partial\lambda'$.

Lemma 3. *Let* $\dot{h}_{t,\lambda}(\theta) := \partial h_t(\theta)/\partial\lambda$ *and* $\dot{h}_{t,\varphi}(\theta) := \partial h_t(\theta)/\partial\varphi$ *be defined in Lemma 2 and write similarly* $\ddot{b}_{t-1,\theta\theta}(\theta) := \partial^2 b_{t-1}(\theta)/\partial\theta\partial\theta'$. *In addition recall that* $\lambda = (w, \alpha, \beta)'$ *and denote further by* $\mu = (w, \alpha)'$. *The following holds:*

$$\ddot{h}_{t,\lambda\lambda}(\theta) = \frac{2\varphi\varepsilon_t^2(\theta)}{h_t(\theta)}\ddot{h}_{t,\lambda}^2(\theta) + \frac{h_t(\theta)}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)}\ddot{b}_{t-1,\lambda\lambda}(\theta),$$

where for i=1,2,3

$$\ddot{b}_{t-1,w\lambda_i}(\theta) = 1 + \sum_{k=0}^{t-1}\beta^{k+1}\sum_{j=0}^{k}\left\{\dot{b}_{t-1-j,\lambda_i}(\theta)\Pi_{t-j}(\theta)\prod_{i=0,i\neq j}^{k}\frac{h_{t-1-i}(\theta)}{h_{t-1-i}(\theta) + \varphi\varepsilon_{t-1-i}^2(\theta)}\right\},$$

$$\ddot{b}_{t-1,\alpha\lambda_i}(\theta) = \sum_{k=0}^{t-1}\beta^k r_{t-1-k}^2\sum_{j=0}^{k-1}\left\{\dot{b}_{t-1-j,\lambda_i}(\theta)\Pi_{t-j}(\theta)\prod_{i=0,i\neq j}^{k-1}\frac{h_{t-1-i}(\theta)}{h_{t-1-i}(\theta) + \varphi\varepsilon_{t-1-i}^2(\theta)}\right\},$$

and for $i = 1, 2$

$$\ddot{b}_{t-1,\beta\mu_i}(\theta) = \sum_{k=0}^{t-1} \beta^k \dot{h}_{t-1-k,\mu_i}(\theta) \prod_{j=0}^{k-1} \frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)} +$$
$$+ \sum_{k=0}^{t-1} \beta^k h_{t-1-k,\mu_i}(\theta) \sum_{j=0}^{k-1} \left\{ \dot{b}_{t-1-j,\mu_i}(\theta)\Pi_{t-j}(\theta) \prod_{i=0, i\neq j}^{k-1} \frac{h_{t-1-i}(\theta)}{h_{t-1-i}(\theta) + \varphi\varepsilon_{t-1-i}^2(\theta)} \right\},$$

and finally

$$\ddot{b}_{t-1,\beta\beta}(\theta) = \sum_{k=0}^{t-1} \beta^k \dot{h}_{t-1-k,\beta}(\theta) \prod_{j=0}^{k-1} \frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)} +$$
$$+ \sum_{k=0}^{t-1} k\beta^{k-1} h_{t-1-k,\beta}(\theta) \prod_{j=0}^{k-1} \frac{h_{t-1-j}(\theta)}{h_{t-1-j}(\theta) + \varphi\varepsilon_{t-1-j}^2(\theta)} +$$
$$+ \sum_{k=0}^{t-1} \beta^k h_{t-1-k,\mu_i}(\theta) \sum_{j=0}^{k-1} \left\{ \dot{b}_{t-1-j,\beta}(\theta)\Pi_{t-j}(\theta) \prod_{i=0, i\neq j}^{k-1} \frac{h_{t-1-i}(\theta)}{h_{t-1-i}(\theta) + \varphi\varepsilon_{t-1-i}^2(\theta)} \right\},$$

where $\Pi_{t-j}(\theta) := \frac{\varphi\varepsilon_{t-j}^2(\theta)}{[h_{t-j}(\theta)+\varphi\varepsilon_{t-j}^2(\theta)]^3} \left(h_{t-j}(\theta) - (h_{t-j}(\theta) + \varphi\varepsilon_{t-j}^2(\theta))^2 \right)$.

Theorem 4. (Asymptotic Normality.) *For the model, given by eq.(2.2)-(2.3), assume Assumptions 1-5 hold. In addition denote by* $\lambda = (w, \alpha, \beta)'$, $\theta = (\lambda', \varphi)'$, *and* $h_t^\star := h_t^\star(\theta_0)$, $\dot{h}_{t,\lambda}^\star(\theta) := \partial h_t^\star(\theta)/\partial\lambda$, $\ddot{h}_{t,\lambda\lambda}^\star(\theta) := \partial^2 h_t^\star(\theta)/\partial\lambda\partial\lambda'$, *and let* $\dot{h}_{t,\lambda}^\star(\theta_0)$ *and* $\ddot{h}_{t,\lambda\lambda}^\star(\theta_0)$ *denote* $\dot{h}_{t,\lambda}^\star(\theta)$ *and* $\ddot{h}_{t,\lambda\lambda}^\star(\theta)$ *evaluated at* λ_0 *respectively, and writing* ε_t^2 *for* $\varepsilon_t^{2,\star}(\theta_0)$. *Then it holds that*

$$\sqrt{T}\left(\widehat{\theta}_T - \theta_0\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathbb{V}_{\theta_0}\right), \tag{2.13}$$

with

$$\mathbb{V}_{\theta_0} = \Sigma_2^{-1}\Sigma_1\Sigma_2^{-1},$$

where

$$\Sigma_1 = \begin{bmatrix} \Sigma_{1,\lambda\lambda} & \Sigma_{1,\lambda\varphi}' \\ \Sigma_{1,\lambda\varphi} & \Sigma_{1,\varphi\varphi} \end{bmatrix},$$

and

$$\Sigma_{1,\lambda\lambda} = \frac{1}{4} E\left[\left\{ \frac{\dot{h}_{t,\lambda}^\star(\theta_0)}{h_t^\star + \varphi_0\varepsilon_t^2}\left(1 - \varepsilon_t^2\right) - \varphi_0 \frac{\dot{h}_{t,\lambda}^\star(\theta_0)}{h_t^\star} \frac{(3\varepsilon_t^2 + \varepsilon_t^4)}{h_t^\star + \varphi_0\varepsilon_t^2} \right\} \left\{ \frac{\dot{h}_{t,\lambda}^\star(\theta_0)}{h_t^\star + \varphi_0\varepsilon_t^2}\left(1 - \varepsilon_t^2\right) - \varphi_0 \frac{\dot{h}_{t,\lambda}^\star(\theta_0)}{h_t^\star} \frac{(3\varepsilon_t^2 + \varepsilon_t^4)}{h_t^\star + \varphi_0\varepsilon_t^2} \right\}' \right]$$

and

$$\Sigma_{1,\varphi\varphi} = \frac{1}{4} E\left[\left\{ \frac{\varepsilon_t^2}{(h_t^\star + \varphi_0\varepsilon_t^2)^2} \left[h_t^\star(\varepsilon_t^2 - 3) - \varphi_0(\varepsilon_t^4 + \varepsilon_t^2) \right] \right\} \left\{ \frac{\varepsilon_t^2}{(h_t^\star + \varphi_0\varepsilon_t^2)^2} \left[h_t^\star(\varepsilon_t^2 - 3) - \varphi_0(\varepsilon_t^4 + \varepsilon_t^2) \right] \right\} \right]$$

and finally $\Sigma_{1,\lambda\varphi} = \Sigma'_{1,\varphi\lambda}$ *is given by*

$$\Sigma_{1,\lambda\varphi} = \frac{1}{4}E\left[\left\{\frac{\dot{h}^{\star}_{t,\lambda}(\theta_0)}{h^{\star}_t + \varphi_0\varepsilon^2_t}\left(1-\varepsilon^2_t\right) - \varphi_0\frac{\dot{h}^{\star}_{t,\lambda}(\theta_0)}{h^{\star}_t}\frac{(3\varepsilon^2_t + \varepsilon^4_t)}{h^{\star}_t + \varphi_0\varepsilon^2_t}\right\}\left\{\frac{\varepsilon^2_t}{(h^{\star}_t + \varphi_0\varepsilon^2_t)^2}\left[h^{\star}_t(\varepsilon^2_t - 3) - \varphi_0(\varepsilon^4_t + \varepsilon^2_t)\right]\right\}\right].$$

Similarly:

$$\Sigma_2 = \begin{bmatrix} \Sigma_{2,\lambda\lambda} & \Sigma'_{2,\lambda\varphi} \\ \Sigma_{2,\lambda\varphi} & \Sigma_{2,\varphi\varphi} \end{bmatrix},$$

where

$$\Sigma_{2,\lambda\lambda} = -\frac{1}{2}E\left\{\frac{h^{2,\star}_t(2\varepsilon^2_t - 1) + \varphi_0\varepsilon^4_t(3\varphi_0 + 4\varphi_0\varepsilon^2_t - 2\varepsilon^2_t) + 2\varphi_0 h^{\star}_t\varepsilon^2_t(5 + 2\varepsilon^2_t)}{(h^{\star}_t + \varphi_0\varepsilon^2_t)^2}\frac{\dot{h}^{\star}_{t,\lambda}(\theta_0)}{h^{\star}_t} + \right.$$
$$\left. + \frac{h^{\star}_t(\varepsilon^2_t - 1) + \varphi_0(3\varepsilon^2_t + \varepsilon^4_t)}{h^{\star}_t + \varphi_0\varepsilon^2_t}\frac{\ddot{h}^{\star}_{t,\lambda\lambda}(\theta_0)}{h^{\star}_t}\right\},$$

and

$$\Sigma_{2,\varphi\varphi} = -\frac{1}{2}E\left[\frac{h^{2,\star}_t\varepsilon^4_t(4\varepsilon^2_t - 13) + 2h^{\star}_t\varepsilon^6_t(2 + 3\varphi_0\varepsilon^2_t) + \varphi_0^2\varepsilon^8_t(1 + 2\varepsilon^2_t)}{(h^{\star}_t + \varphi_0\varepsilon^2_t)^4}\right],$$

and finally $\Sigma_{2\lambda,\varphi} = \Sigma'_{2,\varphi\lambda}$ *is given by*

$$\Sigma_{2,\varphi\lambda} = -\frac{1}{2}E\left[\frac{\dot{h}^{\star}_{t,\lambda}(\theta_0)}{h^{\star}_t}\frac{h^{2,\star}_t\varepsilon^2_t(6 - \varepsilon^2_t) + 2\varphi_0 h^{\star}_t\varepsilon^4_t(2\varepsilon^2_t - 3) + \varphi_0^2\varepsilon^6_t(4 + 5\varepsilon^2_t)}{(h^{\star}_t + \varphi_0\varepsilon^2_t)^3}\right],$$

and $\dot{h}_{t,\lambda}(\theta)$ *and* $\dot{h}_{t,\varphi}(\theta)$ *are defined in Lemma 2 and* $\ddot{h}_{t,\lambda\lambda}(\theta)$ *are defined in Lemma 3 .*

Remark 3. Whenever $\varepsilon_t \sim \mathcal{N}(0,1)$, it then further follows that $\Sigma_1 = \Sigma_2$. In addition, notice that the GARCH model is a nested model, and therefore whenever $\varphi_0 = 0$, the familiar result is obtained for $\theta_0 = (w_0, \alpha_0, \beta_0, 0)$:

$$\Sigma_1 = \frac{1}{2}E\left[\frac{\dot{h}^{\star}_{t,\theta}(\theta_0)\dot{h}^{\star}_{t,\theta}(\theta_0)}{h^{2,\star}_t}\right] \quad \text{and} \quad \Sigma_2 = \frac{1}{4}\left(E\left[\varepsilon^4_t\right] - 1\right)E\left[\frac{\dot{h}^{\theta}_{t,\theta}(\theta_0)\dot{h}^{\star}_{t,\theta}(\theta_0)}{h^{2,\star}_t}\right], \quad h^{\star}_t = b^{\star}_{t-1}.$$

2.3 Simulations

To investigate the large sample behaviour of the QMLE estimates I simulate $N = 500$ sets of data from the model described by eq. (2.2)-(2.3) and compute the maximum likelihood estimates (for each of the parameters) for each simulation. I set the parameter vector to $\theta_0 = (0.03, 0.91, 0.03, 0.03)'$ and consider samples of the size $T = 1000,\ 5000$ and 10000. In addition, I do simulations for $\varepsilon_t \sim \mathcal{N}(0,1)$ and $\varepsilon_t \sim t_{15}$. The simulation results are presented in the Table 1 (Appendix B) and suggest that biases and the size of errors generally decrease as the sample size T increases, suggesting consistency. Moreover, the QQ-plots for each set of the simulations suggest that the distribution of the estimates is very close to normal distribution. This is especially true if $\varepsilon_t \sim \mathcal{N}(0,1)$ and is a reasonable approximation when $\varepsilon_t \sim t_{15}(t_{35})$.

2.4 Conclusion

This chapter develops an asymptotic theory for the QMLE of the Real-time GARCH model presented in Chapter 1. In particular, I prove strong consistency and asymptotic normality of the parameter vector. These results can be thought of as a generalisation of the existing asymptotic theory for the QMLE for the standard GARCH(1,1) model. I also show in my simulations that consistency and asymptotic normality holds for reasonable sample sizes and different specifications of the error term. I believe that the i.i.d. assumption on the error term can be relaxed, although some proofs would have to be modified. In addition, I believe that the results of this paper can be generalised to the nonstationary case where $\alpha + \beta \geq 1$, which would be an extension of the results of Jensen and Rahbek (2004). I leave this for the future research.

References

Andersen, T., Bollerslev, T., Diebold, F.X. and Labys, P., 2003, Modeling and forecasting realized volatility, *Econometrica* 71, 529-626.

Bougerol, P., 1993, Kalman filtering with random coefficients and contractions, *SIAM J. Control Optim* 31, 942-959.

Bougerol, P., and Picard, N., 1992, Strict stationarity of generalized autoregressive processes, *The Annals of Probability*, 20, 1714-1730.

Bougerol, P., and Picard, N., 1992, Stationarity of GARCH processes and some nonnegative time series, *Journal of Econometrics*, 52, 115-127.

Buraczewski, D., Damek, E., and Mikosch, T, 2016, Stochastic models with power-law tails. The equation $X = AX + B$, Springer.

Francq C., and Zakoïan J.M., 2010, GARCH models: structure, statistical inference and Financial applications, Wiley.

Francq C., and Zakoïan J.M., 2004, Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes, *Bernoulli* 10, 605-637.

Francq C., and Zakoïan J.M., 2007, Quasi-maximum likelihood estimation in GARCH processes when some coefficients are equal to zero, *Stochastic Processes and their Applications* 117 (9), 1265-1284.

Hall, P and C. Heyde, 1980, Martingale limit theory and its application. New York: Academic Press.

Hansen, P.R., Huang, Z. and Shek, H.H., 2011, Realized GARCH: a joint model for returns and realized measures of volatility, *Journal of Applied Econometrics* 27, 877-906.

Jensen S.T., and Rahbek A., 2004, Asymptotic inference for nonstationary GARCH, *Econometric Theory* 20, 1203-1226.

Kristensen, D., and Han, H., 2014, Asymptotic theory for the QMLE in GARCH-X models with stationary and nonstationary covariates, *Journal of Business & Economic Statistics*, 32 (3), 416 - 429.

Kristensen, D., and Rahbek, A., 2009, Asymptotics of the QMLE for non-linear ARCH models, *Journal of Time Series Econometrics* 1 (1), Article 2.

Lee, S.-W. and Hansen, B.E., 1994, Asymptotic theory for the GARCH (1,1) quasi- maximum likelihood estimator, *Econometric Theory* 10, 29-52.

Lumsdaine R. L., 1996, Consistency and asymptotic normality of the quasi-maximum likelihood estimator in IGARCH(1,1) and covariance stationary GARCH(1,1) models, *Econometrica* 64, 575596.

Pollard, D., 1984, *Convergence of stochastic processes.* Springer, New York.

Straumann, D., 2005, Estimation in conditionally heteroskedastic time series models, Lecture Notes in Statistics 181. Springer, Berlin.

2.5 Appendix A.

Proof of Theorem 1.

To prove Theorem 1 I appeal to the Theorem 2.1.3 in Buraczewski et. al. (2016) to establish stationarity and ergodicity of the joint process. The joint process (r_t^2, h_t) can be written as the nonnegative stochastic recurrence equation, see eq.(2.10). Then under Assumption 1, condition 2 in Theorem 2.1.3 in Buraczewski et. al. (2016) holds, which allows me to conclude that there exists a.s.-unique causal ergodic strictly stationary solution to the recurrence equation (2.10). ■

Proof of Lemma 1. I first consider $h_t(\theta)$:

$$h_t(\theta) = w + \alpha r_{t-1}^2 + \beta h_{t-1}(\theta) + \varphi \varepsilon_t^2(\theta) = w + \varphi \varepsilon_t^2(\theta) + \alpha r_{t-1}^2 + w\beta + \varphi\beta\varepsilon_{t-1}^2(\theta) + \beta^2 h_{t-2}(\theta) =$$
$$= \cdots = w\sum_{j=0}^{t}\beta^j + \varphi\sum_{j=0}^{t}\beta^j\varepsilon_{t-j}^2(\theta) + \alpha\sum_{j=0}^{t-1}\beta^j r_{t-1-j}^2 + \beta^t h_0(\theta).$$

Therefore, provided that $\beta < \overline{\beta} < 1$ and the initial values $h_0(\theta) < \infty$ and $h_0^\star(\theta) < \infty$, and assuming that the high-level Assumption 2 holds, it then follows:

$$\sup_{\theta\in\Theta}|h_t(\theta) - h_t^\star(\theta)| = \sup_{\theta\in\Theta}\left|\varphi\sum_{j=0}^{t}\beta^j\left(\varepsilon_{t-j}^2(\theta) - \varepsilon_{t-j}^{2,\star}(\theta)\right) + \beta^t\left(h_0(\theta) - h_0^\star(\theta)\right)\right| \leq$$
$$\leq \sum_{j=0}^{t}\sup_{\theta\in\Theta}\left|\varphi\beta^j\left(\varepsilon_{t-j}^2(\theta) - \varepsilon_{t-j}^{2,\star}(\theta)\right)\right| + \sup_{\theta\in\Theta}\left|\beta^t(h_0(\theta) - h_0^\star(\theta))\right| \leq$$
$$\leq \overline{\varphi}\sup_{\theta\in\Theta}\left|\varepsilon_t^2(\theta) - \varepsilon_t^{2,\star}(\theta)\right|\sum_{j=0}^{t}\overline{\beta}^j + \sup_{\theta\in\Theta}\left|\beta^t(h_0(\theta) - h_0^\star(\theta))\right| \leq \overline{\varphi}K\rho^t\sum_{j=0}^{t}\overline{\beta}^j + \overline{\beta}^t C_1 =$$
$$= \frac{\overline{\varphi}}{1-\overline{\beta}}K\rho^t(1-\overline{\beta}^t) + \overline{\beta}^t C_1, \tag{2.14}$$

where $C_1 = \sup_{\theta\in\Theta}|h_0(\theta) - h_0^\star(\theta)| < \infty$. Using this we can further derive:

$$\sup_{\theta\in\Theta}\left|L_T(\theta)-L_T^\star(\theta)\right| \le T^{-1}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\left\{\frac{1}{2}\left|\frac{h_t^\star(\theta)-h_t(\theta)}{h_t^\star(\theta)h_t(\theta)}\right|r_t^2+\frac{1}{2}\left|\log\left(\frac{h_t^\star(\theta)}{h_t(\theta)}\right)\right|+\left|\log\left(\frac{h_t(\theta)+\varphi\varepsilon_t^2(\theta)}{h_t^\star(\theta)+\varphi\varepsilon_t^{2,\star}(\theta)}\right)\right|\right\}$$

$$\le T^{-1}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\left\{\frac{1}{2}\left|\frac{h_t^\star(\theta)-h_t(\theta)}{h_t^\star(\theta)h_t(\theta)}\right|r_t^2+\frac{1}{2w}\left|h_t^\star(\theta)-h_t(\theta)\right|+\frac{1}{w}\left(\left|h_t(\theta)-h_t^\star(\theta)\right|+\varphi\left|\varepsilon_t^2(\theta)-\varepsilon_t^{2,\star}(\theta)\right|\right)\right\}$$

$$\le \frac{1}{2\underline{w}^2}T^{-1}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\left|h_t(\theta)-h_t^\star(\theta)\right|r_t^2+\frac{3}{2\underline{w}}T^{-1}\sum_{t=1}^{T}\sup_{\theta\in\Theta}\left|h_t(\theta)-h_t^\star(\theta)\right|+\overline{\varphi}T^{-1}\sum_{t=1}^{T}\sup_{\theta in\Theta}\left|\varepsilon_t^2(\theta)-\varepsilon_t^{2,\star}(\theta)\right|\le$$

$$\le T^{-1}\frac{1}{2\underline{w}^2}\frac{\overline{\varphi}}{1-\overline{\beta}}K\sum_{t=1}^{T}\rho^t(1-\overline{\beta}^t)r_t^2+T^{-1}\frac{1}{2\underline{w}^2}C_1\sum_{t=1}^{T}\overline{\beta}^t r_t^2+T^{-1}\frac{3}{2\underline{w}}\frac{\overline{\varphi}}{1-\overline{\beta}}K\sum_{t=1}^{T}\rho^t(1-\overline{\beta}^t)+$$

$$+T^{-1}\frac{3}{2\underline{w}}C_1\sum_{t=1}^{T}\overline{\beta}^t+\overline{\varphi}KT^{-1}\sum_{t=1}^{T}\rho^t. \tag{2.15}$$

where in the second and third lines of the derivation above I used the fact that for $x,y>0$, $\left|\log\frac{x}{y}\right|\le\frac{|x-y|}{min(x,y)}$. By applying Borel-Cantelli lemma, existence of moments of orders $s>0$ (by Assumption 2) for r_t^2 and Markov inequality we further can conclude that:

$$\sum_{t=1}^{\infty}P\left(\rho^t r_t^2>\epsilon\right)\le\sum_{t=1}^{\infty}\frac{E\left[r_t^{2s}\right]}{\epsilon^s}<\infty,$$

Similar calculations hold to show that $(\rho\overline{\beta})^t r_t^2\to 0$ almost surely and $\overline{\beta}^t r_t^2\to 0$ almost surely. In addition, it also holds that

$$\lim_{T\to\infty}\sum_{t=1}^{T}\overline{\beta}^t=\frac{1}{1-\overline{\beta}}<\infty\qquad\text{and}\qquad\lim_{T\to\infty}\sum_{t=1}^{T}\rho^t=\frac{1}{1-\rho}<\infty,$$

and similarly

$$\lim_{T\to\infty}\sum_{t=1}^{T}\rho^t\frac{1-\overline{\beta}^t}{1-\overline{\beta}}=\frac{1}{1-\overline{\beta}}\lim_{T\to\infty}\left(\sum_{t=1}^{T}\rho^t-\sum_{t=1}^{T}(\rho\overline{\beta})^t\right)=\frac{1}{1-\overline{\beta}}\left(\frac{\rho}{1-\rho}-\frac{\overline{\beta}\rho}{1-\overline{\beta}\rho}\right)<\infty.$$

Combining all of the above allows now to conclude that $\sup_{\theta\in\Theta}\left|L_T(\theta)-L_T^\star(\theta)\right|=o_p(1/T)$, which concludes the proof of Lemma 1. ■

Proof of Theorem 2.

In light of Lemma 1 we now work with the stationary sequences. In particular let $\widetilde{L}_T^\star(\theta)$ and $L_T^\star(\theta)$ denote two stationary versions of $L_T(\theta)$ that differ with their initial values. The proof of Theorem 2 closely follows Francq and Zakoïan (2004), Theorem 7.1 for the standard GARCH(1,1) case by adapting it to my case. More precisely, in order to establish the result in Theorem 2, I will need to show the following intermediate results:

(a) $\lim_{T\to\infty} \sup_{\theta\in\Theta} |L^{\star}_T(\theta) - \widetilde{L}^{\star}_T(\theta)| = 0$, a.s. where

(b) $(\exists t \in \mathbb{Z}$ such that $h^{\star}_t(\theta) = h^{\star}_t(\theta_0),\ P_{\theta_0}$ a.s.$) \Rightarrow \theta = \theta_0$.

(c) $E_{\theta_0}|l^{\star}_t(\theta_0)| < \infty$, and if $\theta \neq \theta_0$, $E_{\theta_0} l^{\star}_t(\theta) > E_{\theta_0} l^{\star}_t(\theta_0)$.

(d) For any $\theta \neq \theta_0$, there exists a neighbourhood $V(\theta)$ such that $\lim\inf_{T\to\infty} \inf_{\theta^{\star}\in V(\theta)} \widetilde{L}^{\star}_T(\theta^{\star}) > E_{\theta_0} l^{\star}_t(\theta_0)$ a.s.

(a) Asymptotic irrelevance of the initial values. The proof is identical to the proof of Lemma 1 for $L^{\star}_T(\theta)$ and $\widetilde{L}^{\star}_T(\theta)$ and therefore is omitted.

(b) Identifiability of the parameter vector. The proof is similar to that of Kristensen and Han (2014) in the GARCH-X case for the identification of the parameter vector. Here and for the rest of the proofs to follow for simplicity I write $\varepsilon^2_t(\theta)$ for $\varepsilon^{2,\star}_t(\theta)$ and ε^2_t for $\varepsilon^{2,\star}_t(\theta_0)$. Via recursive substitution:

$$h^{\star}_t(\theta) = w+\alpha r^2_{t-1}+\beta h^{\star}_{t-1}(\theta)+\varphi\varepsilon^2_t(\theta) = w+\alpha r^2_{t-1}+\varphi\varepsilon^2_t+\beta\left[w + \alpha r^2_{t-2} + \beta h^{\star}_{t-2}(\theta) + \varphi\varepsilon^2_{t-1}(\theta)\right] =$$
$$= w + \beta w + \alpha r^2_{t-1} + \alpha\beta r^2_{t-2} + \varphi\varepsilon^2_t + \varphi\beta\varepsilon^2_{t-1} + \beta^2 h^{\star}_{t-3}(\theta) = \cdots =$$
$$= w\sum_{j=1}^{\infty}\beta^{j-1} + \alpha\sum_{j=1}^{\infty}\beta^{j-1}r^2_{t-j} + \varphi\sum_{j=1}^{\infty}\beta^{j-1}\varepsilon^2_{t+1-j}(\theta).$$

Denote by $\nu_j := (\alpha\beta^{j-1}, \varphi\beta^{j-1})$. I now want to establish that if $h^{\star}_t(\theta) = h^{\star}_t(\theta_0),\ P_{\theta_0}$ a.s. then $w = w_0$ and $\nu_j(\theta) = \nu_j(\theta_0)$ and therefore $\theta = \theta_0$. I prove this by contradiction. Assume $h^{\star}_t(\theta) = h^{\star}_t(\theta_0)$ and let $m > 0$ to be the smallest integer for which $h^{\star}_t(\theta) = h^{\star}_t(\theta_0)$ yet $\nu_j(\theta) \neq \nu_j(\theta_0)$ (whenever $\nu_j(\theta) = \nu_j(\theta_0)$ it follows that $w = w_0$). Then from $h^{\star}_t(\theta) = h^{\star}_t(\theta_0)$:

$$a_0 r^2_{t-m} + \varphi_0\beta_0^{-1}\varepsilon^2_{t-m+1} - \varphi\beta^{-1}\varepsilon_{t-m+1}(\theta) - =$$
$$= w - w_0 + \sum_{j=1}^{\infty} a_j r^2_{t-m-j} + \sum_{j=1}^{\infty}\left(\varphi_0\beta_0^{j-1}\varepsilon^2_{t-m-j} - \varphi\beta^{j-1}\varepsilon^2_t(\theta)\right), \quad (2.16)$$

where $a_j := \alpha_0\beta_0^{j-1} - \alpha\beta^{j-1}$ and $b_j := \varphi_0\beta_0^{j-1} - \varphi\beta^{j-1}$. The right-hand side of (2.16) belongs to $\mathcal{F}_{t-m-1}$, where $\mathcal{F}_{t-m-1}$ is the information set up to and including time $t-m-1$. Therefore conditional on $\mathcal{F}_{t-m-1}$ the right-hand side of (2.16) is constant. This implies that the left-hand side of (2.16) is constant, which however is ruled out by Assumption 3(ii).

(c) The limit criterion is minimised at the true value. Using the conventional notation

$x^- = \max(-x, 0)$ and $x^+ = \max(x, 0)$ and writing again ε_t^2 for $\varepsilon_t^{\star,2}(\theta)$ one gets:

$$E_{\theta_0} l_t^{\star,-}(\theta) \leq E_{\theta_0}\left\{\frac{1}{2}\frac{r_t^2}{h_t^\star(\theta)} + \frac{1}{2}\log^- h_t^\star(\theta) + \log^-\left(1 + \frac{\varphi\varepsilon_t^2}{h_t^\star(\theta)}\right)\right\} \leq$$
$$\leq E_{\theta_0}\log^- h_t^\star(\theta) \leq \max\{0, -\log w\} < \infty.$$

Moreover, Assumption 1 guarantees the existence of moments of order $s > 0$ for r_t^2 and $h_t(\theta)$. Using Jensen's inequality, I can write

$$E_{\theta_0}\log^+ h_t^\star(\theta_0) < \infty$$

as

$$E_{\theta_0}\log h_t^\star(\theta_0) = E_{\theta_0}\frac{1}{s}\log\{h_t^\star(\theta_0)\}^s \leq \frac{1}{s}\log E_{\theta_0}\{h_t^\star(\theta_0)\}^s < \infty. \tag{2.17}$$

Then it follows that

$$E_{\theta_0} l_t^\star(\theta_0) = E_{\theta_0}\left\{\frac{1}{2}\log(2\pi) + \frac{1}{2}\varepsilon_t^2 - \log\left(\frac{\sqrt{h_t^\star(\theta_0)}}{h_t^\star(\theta_0) + \varphi_0\varepsilon_t^2}\right)\right\} =$$
$$E_{\theta_0}\left\{\frac{1}{2}\log(2\pi) + \frac{1}{2}\varepsilon_t^2 - \frac{1}{2}\log h_t^\star(\theta_0) + \log\left(1 + \frac{\varphi_0\varepsilon_t^2}{h_t^\star(\theta_0)}\right)\right\} \leq$$
$$\leq \frac{1}{2} - \frac{1}{2}\underbrace{E_{\theta_0}\log h_t^\star(\theta_0)}_{<\infty \text{ by } (2.17)} + \underbrace{E_{\theta_0}\left(\frac{\varphi_0\varepsilon_t^2}{b_{t-1}^\star(\theta_0) + \varphi_0\varepsilon_t^2}\right)}_{<\infty} < \infty.$$

Since I already showed that $E_{\theta_0}\log l_t^{\star,-}(\theta_0) < \infty$, it then follows that $E_{\theta_0} l_t^\star(\theta_0)$ is well defined in $\mathbb{R}$. Using again the fact that for $x > 0$, $x - 1 \geq \log x$ (with equality if and only if x=1) and Jensen's inequality I in addition have:

$$E_{\theta_0} l_t^\star(\theta) - E_{\theta_0} l_t^\star(\theta_0) = \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \frac{1}{2}E_{\theta_0}\frac{h_t^\star(\theta)\varepsilon_t^2}{h_t^\star(\theta_0)} - \frac{1}{2}E_{\theta_0}\varepsilon_t^2 + E_{\theta_0}\log\left(\frac{1 + \frac{\varphi\varepsilon_t^2}{h_t^\star(\theta)}}{1 + \frac{\varphi_0\varepsilon_t^2}{h_t^\star(\theta_0)}}\right) \geq$$
$$\geq \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \frac{1}{2}\left[E_{\theta_0}\frac{h_t^\star(\theta)\varepsilon_t^2}{h_t^\star(\theta_0)} - 1\right] = \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \frac{1}{2}\left[E_{\theta_0}\frac{h_t^\star(\theta)\varepsilon_t^2}{h_t^\star(\theta_0)} - 1\right] \geq$$
$$\geq \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \frac{1}{2}\log E_{\theta_0}\frac{h_t^\star(\theta)\varepsilon_t^2}{h_t^\star(\theta_0)} \geq \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \frac{1}{2}E_{\theta_0}\log\frac{h_t^\star(\theta)\varepsilon_t^2}{h_t^\star(\theta_0)} \geq$$
$$\geq \frac{1}{2}E_{\theta_0}\left\{\log\frac{h_t^\star(\theta_0)}{h_t^\star(\theta)} + \log\frac{h_t^\star(\theta)}{h_t^\star(\theta_0)}\right\} = 0$$

with equality if and only if $h_t^\star(\theta) = h_t^\star(\theta_0)$, which in the view of (b) is equivalent to the condition of $\theta = \theta_0$.

(d) Compactness of Θ and ergodicity of the score function. Ergodicity of the score function follows directly from Theorem 1 and the fact that the score function is a well-defined measurable function of the joint process $(r_t^2, h_t^\star)$.

The proof of Theorem 2 is then follows by combining a)-d) in conjunction with Lemma 1 and appealing to the LLN for the stationary and ergodic sequences. ■

Proof of Lemma 2. Recall that $h_t(\theta)$ can be written as follows:

$$h_t = b_{t-1}(\theta) + \varphi\varepsilon_t^2(\theta). \tag{2.18}$$

Differentiating eq.(2.18) with respect to λ I get:

$$\dot{h}_{t,\lambda}(\theta) = \dot{b}_{t-1,\lambda}(\theta) + 2\varphi\varepsilon_t(\theta)\dot{\varepsilon}_{t,\lambda}(\theta) = \dot{b}_{t-1,\lambda}(\theta) - \varphi\varepsilon_t^2(\theta)\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)}.$$

Rearranging I get:

$$\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} = \frac{\dot{b}_{t-1,\lambda}(\theta)}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)}. \tag{2.19}$$

Differentiating eq.(2.18) with respect to φ I get:

$$\dot{h}_{t,\lambda}(\theta) = \dot{b}_{t-1,\lambda}(\theta) + \varepsilon_t^2(\theta) + 2\varphi\varepsilon_t(\theta)\dot{\varepsilon}_{t,\varphi}(\theta) = \varepsilon_t^2(\theta) - \varepsilon_t^2(\theta)\varphi\frac{1}{h_t(\theta)}\dot{h}_{t,\varphi}(\theta).$$

Rearranging I further get:

$$\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)} = \frac{\varepsilon_t^2(\theta)}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)}. \tag{2.20}$$

I next derive the derivatives of $b_{t-1}(\theta)$ with respect to each of the elements of the parameter vector θ. I first get via recursive substitution:

$$b_{t-1}(\theta) = w+\alpha r_{t-1}^2+\beta h_{t-1}(\theta) = w+\alpha r_{t-1}^2+\beta\left[b_{t-2}(\theta) + \varphi\varepsilon_{t-1}^2(\theta)\right] = w+\alpha r_{t-1}^2+\beta b_{t-2}(\theta)+\beta\varphi\varepsilon_{t-1}^2(\theta).$$

Taking the derivative with respect to w:

$$\dot{b}_{t-1,w}(\theta) = 1 + \beta\dot{b}_{t-2,w}(\theta) + \beta\varphi 2\varepsilon_{t-1}(\theta)\dot{\varepsilon}_{t,w}(\theta)$$

In addition,

$$\dot{\varepsilon}_{t,\theta}(\theta) = -\frac{1}{2}\varepsilon_t(\theta)\frac{\dot{h}_{t,\theta}(\theta)}{h_t(\theta)},$$

and

$$\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} = \frac{\dot{b}_{t-1,\lambda}(\theta)}{b_{t-1}(\theta) + 2\varphi\varepsilon_t^2(\theta)} \quad \text{and} \quad \frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)} = \frac{\varepsilon_t^2(\theta)}{b_{t-1}(\theta) + 2\varphi\varepsilon_t^2(\theta)}.$$

Therefore,

$$\dot{b}_{t-1,w}(\theta) = 1+\beta\dot{b}_{t-2,w}(\theta)-\beta\varphi\frac{\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\dot{b}_{t-2,w}(\theta) = 1+\beta\left[\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right]\dot{b}_{t-2,w}(\theta) =$$

$$= 1+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left\{1+\beta\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,w}(\theta)\right\} =$$

$$1+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)+\beta^2\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,w}(\theta) =$$

$$\cdots = 1+\sum_{k=0}^{t-1}\beta^{k+1}\prod_{j=0}^{k}\frac{b_{t-2-j}(\theta)+\varphi\varepsilon^2_{t-1-j}(\theta)}{b_{t-2-j}(\theta)+2\varphi\varepsilon^2_{t-1-j}(\theta)}.$$

I now calculate the derivative of $b_{t-1}(\theta)$ with respect to α:

$$\dot{b}_{t-1,\alpha}(\theta) = r^2_{t-1}+\beta\dot{b}_{t-2,\alpha}(\theta)-\beta\varphi\frac{\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\dot{b}_{t-2,\alpha}(\theta) = r^2_{t-1}+$$

$$+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\dot{b}_{t-2,\alpha}(\theta) = r^2_{t-1}+$$

$$+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left[r^2_{t-2}+\beta\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,\alpha}(\theta)\right] =$$

$$= r^2_{t-1}+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)r^2_{t-2}+\beta^2\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,\alpha}(\theta) =$$

$$= \cdots = \sum_{k=0}^{t-1}\beta^k r^2_{t-1-k}\prod_{j=0}^{k-1}\frac{b_{t-2-j}(\theta)+\varphi\varepsilon^2_{t-1-j}(\theta)}{b_{t-2-j}(\theta)+2\varphi\varepsilon^2_{t-1-j}(\theta)}.$$

I now calculate the derivative of $b_{t-1}(\theta)$ with respect to β:

$$\dot{b}_{t-1,\beta}(\theta) = b_{t-2}(\theta)+\beta\dot{b}_{t-2,\beta}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)-\beta\varphi\frac{\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\dot{b}_{t-2,\beta}(\theta) =$$

$$= h_{t-1}(\theta)+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left(h_{t-2}(\theta)+\beta\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,\beta}(\theta)\right) =$$

$$= h_{t-1}(\theta)+\beta\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)h_{t-2}(\theta)+\beta^2\left(\frac{b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left(\frac{b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right)\dot{b}_{t-3,\beta}(\theta) =$$

$$= \sum_{k=0}^{t-1}\beta^k h_{t-1-k}(\theta)\prod_{j=0}^{k-1}\frac{b_{t-2-j}(\theta)+\varphi\varepsilon^2_{t-1-j}(\theta)}{b_{t-2-j}(\theta)+2\varphi\varepsilon^2_{t-1-j}(\theta)}.$$

Finally, I differentiate $b_{t-1}(\theta)$ with respect to φ:

$$
\begin{aligned}
\dot{b}_{t-1,\varphi}(\theta) &= \beta\dot{b}_{t-2,\varphi}(\theta) + \beta\varepsilon^2_{t-1}(\theta) - \beta\varphi\frac{\varepsilon^4_{t-1}(\theta)}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\dot{b}_{t-2,\varphi}(\theta) = \\
&= \beta\left(\frac{\varepsilon^2_{t-1}(\theta)\left[b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)\right]}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right) + \beta\dot{b}_{t-2,\varphi}(\theta) = \\
&= \beta\left(\frac{\varepsilon^2_{t-1}(\theta)\left[b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)\right]}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right) + \beta^2\left(\frac{\varepsilon^2_{t-1}(\theta)\left[b_{t-2}(\theta)+\varphi\varepsilon^2_{t-1}(\theta)\right]}{b_{t-2}(\theta)+2\varphi\varepsilon^2_{t-1}(\theta)}\right)\left(\frac{\varepsilon^2_{t-2}(\theta)\left[b_{t-3}(\theta)+\varphi\varepsilon^2_{t-2}(\theta)\right]}{b_{t-3}(\theta)+2\varphi\varepsilon^2_{t-2}(\theta)}\right) + \\
&+ \beta^2\dot{b}_{t-3,\varphi}(\theta) = \cdots = \sum_{k=1}^{t-1}\beta^k\prod_{j=1}^{k}\frac{\varepsilon^2_{t-j}(\theta)\left[b_{t-1-j}(\theta)+\varphi\varepsilon^2_{t-j}(\theta)\right]}{b_{t-1-j}(\theta)+2\varphi\varepsilon^2_{t-j}(\theta)}.
\end{aligned}
$$

This completes the proof of Lemma 2. ■

Proof of Lemma 3. From Lemma 2 we get:

$$
\begin{aligned}
\ddot{h}_{t,\lambda\lambda}(\theta) &= \dot{b}_{t-1,\lambda}(\theta)\left[\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)+\varphi\varepsilon^2_t(\theta)} - \frac{h_t(\theta)}{(h_t(\theta)+\varphi\varepsilon^2_t(\theta))^2}\left(\dot{h}_{t,\lambda}(\theta)+2\varphi\varepsilon^2_t(\theta)\dot{\varepsilon}_{t,\lambda}(\theta)\right)\right] + \\
&+ \frac{h_t\theta}{h_t(\theta)+\varphi\varepsilon^2_t(\theta)}\ddot{b}_{t-1,\lambda}(\theta) = \dot{b}_{t-1,\lambda}(\theta)\left[\frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)+\varphi\varepsilon^2_t(\theta)} - \frac{h_t(\theta)}{(h_t(\theta)+\varphi\varepsilon^2_t(\theta))^2}\dot{h}_{t,\lambda}(\theta)\left(1-\frac{\varphi\varepsilon^2_t(\theta)}{h_t(\theta)}\right)\right] + \\
&+ \frac{h_t\theta}{h_t(\theta)+\varphi\varepsilon^2_t(\theta)}\ddot{b}_{t-1,\lambda}(\theta) = \frac{2\varepsilon^2_t(\theta)}{h_t(\theta)}\dot{h}^2_{t,\lambda} + \frac{h_t(\theta)}{h_t(\theta)+\varphi\varepsilon^2_t(\theta)}\ddot{b}_{t-1,\lambda\lambda},
\end{aligned}
$$

where $\ddot{b}_{t-1,\lambda\lambda}(\theta)$ is obtained by direct differentiation of $\dot{b}_{t-1,\lambda}(\theta)$ in Lemma 2. ■

Proof of Theorem 3. I first derive the score function. Since it is not possible to write down the score function as a function of the whole parameter vector $\theta = (w, \alpha, \beta, \varphi)$, I will separately write it down for the parameter vector $\lambda := (w, \alpha, \beta)$ and for φ. In order to proceed I will be using Lemmas 4 and 5 that I state below.

Lemma 4. *Consider the following indefinite integral*

$$
\int \frac{\xi_1 x^4 + \xi_2 x^2 + \xi_3}{(x^2+\xi_4)^2} e^{-x^2/2} dx. \tag{2.21}
$$

If the ξ_1, ξ_2, ξ_3 and ξ_4 are related as follows:

$$
\xi_3 = \xi_1 - \xi_2 \quad \textit{and} \quad \xi_4 = (\xi_2 - \xi_1)/\xi_1,
$$

then it holds that:

$$\int \frac{\xi_1 x^4 + \xi_2 x^2 + \xi_3}{(x^2+\xi_4)^2} e^{-x^2/2} dx = -\frac{\xi_1^2 e^{-x^2/2} x}{\xi_1(x^2-1)+\xi_2}. \tag{2.22}$$

The proof follows by direct differentiation of eq.(2.22). ■

Lemma 5. *Consider the following indefinite integral*

$$\int \frac{\chi_1 x^6 + \chi_2 x^4 + \chi_3 x^2}{(x^2+\chi_4)^2} e^{-x^2/2} dx. \tag{2.23}$$

If the χ_1, χ_2, χ_3 and χ_4 are related as follows:

$$\xi_3 = -3(\chi_1+\chi_2) \quad \textit{and} \quad \chi_4 = (\chi_2+\chi_1)/\chi_1,$$

then it holds that:

$$\int \frac{\chi_1 x^6 + \chi_2 x^4 + \chi_3 x^2}{(x^2+\chi_4)^2} e^{-x^2/2} dx = -\frac{1}{2}\frac{e^{-x^2/2} x^3}{(x^2+\chi_4)}. \tag{2.24}$$

The proof follows by direct differentiation of eq.(2.24). ■

I next proceed by deriving the score function separately for λ and for φ. I start with λ first. Note that using eq.(2.2)-(2.3) the conditional density function can be written as follows:

$$f_r(r|\mathcal{F}_{t-1}) = \frac{r}{\varepsilon_t(\theta)\sqrt{b_{t-1}^2(\theta)+4r^2\varphi}} f_\epsilon(\varepsilon_t(\theta)).$$

First observe that from eq.(2.6) it holds that

$$\frac{r_t}{\varepsilon_t(\theta)\sqrt{b_{t-1}^2(\theta)+4r^2\varphi}} = \frac{\partial \varepsilon_t(\theta)}{\partial r_t} = \frac{\sqrt{h_t(\theta)}}{h_t(\theta)+\varphi\varepsilon_t^2(\theta)}. \tag{2.25}$$

I therefore can rewrite the conditional density as follows:

$$f_r(r|\mathcal{F}_{t-1}) = \frac{\partial \varepsilon_t(\theta)}{\partial r} f_\epsilon(\varepsilon_t(\theta)).$$

I consider the likelihood estimators based on the minimisation of the negative log-likelihood, given by

$$L_T(\theta) = -\frac{1}{T}\sum_{t=1}^{T} l_t(\theta),$$

where

$$l_t(r_t|\theta,\mathcal{F}_{t-1}) = \log\left[\frac{\partial \varepsilon_t(\theta)}{\partial r_t}\right] + \log f_\epsilon(\varepsilon_t(\theta)) = \log\left[\frac{\partial \varepsilon_t(\theta)}{\partial r_t}\right] - \frac{1}{2}\log(2\pi) - \frac{1}{2}\varepsilon_t^2(\theta).$$

The first derivative of $l_t(\theta)$ with respect to a generic θ is given by:

$$\frac{\partial l_t(\theta)}{\partial \theta} = -\left(\left[\frac{\partial \varepsilon_t\left(\theta\right)}{\partial r_t} \right]^{-1} \frac{\partial^2 \varepsilon_t\left(\theta\right)}{\partial r_t \partial \theta} - \varepsilon_t\left(\theta\right) \frac{\partial \varepsilon_t(\theta)}{\partial \theta} \right). \tag{2.26}$$

From eq.(2.25) it holds that

$$\frac{\partial \varepsilon_t\left(\theta\right)}{\partial r_t} = h_t(\theta)^{-1/2} - \frac{1}{2} h_t(\theta)^{-3/2} r_t \frac{\partial h_t(\theta)}{\partial r_t} = h_t(\theta)^{-1/2} - \varphi \varepsilon_t^2\left(\theta\right) \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial r_t} = \frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi \varepsilon_t^2(\theta)} \tag{2.27}$$

and

$$\dot{\varepsilon}_{t,\theta}(\theta) = -\frac{1}{2} r_t h_t(\theta)^{-3/2} \dot{h}_{t,\theta}(\theta) = -\frac{1}{2} \frac{\varepsilon_t\left(\theta\right)}{h_t(\theta)} \dot{h}_{t,\theta}(\theta). \tag{2.28}$$

The above equation holds for the whole parameter vector $\theta = (\lambda', \varphi)'$. Differentiating eq.(2.27) with respect to λ I get

$$\begin{aligned}
\frac{\partial^2 \varepsilon_t\left(\theta\right)}{\partial r_t \partial \lambda} &= \frac{1}{2} h_t(\theta)^{-1/2} \dot{h}_{t,\lambda}(\theta) \left[h_t(\theta) + \varphi \varepsilon_t^2\left(\theta\right) \right]^{-1} - \\
&- \left(h_t(\theta) + \varphi \varepsilon_t^2\left(\theta\right) \right)^{-2} \left[\dot{h}_{t,\lambda}(\theta) + 2\varphi \varepsilon_t\left(\theta\right) \dot{\varepsilon}_{t,\lambda}(\theta) \right] h_t(\theta)^{1/2} = \frac{1}{2} \frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi \varepsilon_t^2(\theta)} \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} - \\
&- \frac{\sqrt{h_t(\theta)}}{(h_t(\theta) + \varphi \varepsilon_t^2(\theta))^2} \left[\dot{h}_{t,\lambda}(\theta) - \varphi \varepsilon_t^2\left(\theta\right) \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} \right] = \frac{1}{2} \frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi \varepsilon_t^2(\theta)} \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} - \\
&- \frac{\sqrt{h_t(\theta)}}{(h_t(\theta) + \varphi \varepsilon_t^2(\theta))^2} \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} \left[h_t(\theta) - \varphi \varepsilon_t^2\left(\theta\right) \right] = \\
&= \frac{\sqrt{h_t(\theta)}}{(h_t(\theta) + \varphi \varepsilon_t^2(\theta))^2} \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} \left[-\frac{1}{2} h_t(\theta) + \frac{3}{2} \varphi \varepsilon_t^2(\theta) \right].
\end{aligned}$$

Plugging the above expressions in eq. (2.26) it then follows:

$$\begin{aligned}
\frac{\partial l_t(\theta)}{\partial \lambda} &= -\left(\left[\frac{\partial \varepsilon_t\left(\theta\right)}{\partial r_t} \right]^{-1} \frac{\partial^2 \varepsilon_t\left(\theta\right)}{\partial r_t \partial \lambda} - \varepsilon_t\left(\theta\right) \frac{\partial \varepsilon_t\left(\theta\right)}{\partial \lambda} \right) = \\
&= \frac{1}{h_t(\theta) + \varphi \varepsilon_t^2(\theta)} \frac{\dot{h}_{t,\lambda}(\theta)}{h_t(\theta)} \left(-\frac{1}{2} h_t(\theta)(1 - \varepsilon_t^2(\theta)) + \frac{1}{2} \varphi (3\varepsilon_t^2(\theta) + \varepsilon_t^4(\theta)) \right) = \\
&= \frac{\dot{b}_{t-1,\lambda}(\theta)}{(b_{t-1}\left(\theta\right) + 2\varphi \varepsilon_t^2(\theta))^2} \left(-\frac{1}{2} b_{t-1}\left(\theta\right) \left(1 - \varepsilon_t^2(\theta)\right) + \varphi \varepsilon_t^4(\theta) + \varphi \varepsilon_t^2(\theta) \right).
\end{aligned}$$

I next derive the score function for the parameter φ.

$$\frac{\partial \varepsilon_t\left(\theta\right)}{\partial r_t} = h_t(\theta)^{-1/2} - \frac{1}{2} h_t(\theta)^{-3/2} r_t \frac{\partial h_t(\theta)}{\partial r_t} = h_t(\theta)^{-1/2} - \frac{1}{2} \varepsilon_t(\theta) \frac{1}{h_t(\theta)} \frac{\partial h_t(\theta)}{\partial r_t}.$$

Using the fact that $\frac{\partial h_t(\theta)}{\partial r_t} = 2\varphi\varepsilon_t(\theta)\frac{\partial \varepsilon_t(\theta)}{\partial r_t}$ and rearranging I get:

$$\frac{\partial \varepsilon_t\left(\theta\right)}{\partial r_t} = \frac{\sqrt{h_t(\theta)}}{h_t(\theta) + \varphi\varepsilon_t^2(\theta)}.$$

In addition, it also holds that

$$\dot{\varepsilon}_{t,\varphi}(\theta) = -\frac{1}{2} r_t h_t(\theta)^{-3/2}\dot{h}_{t,\varphi}(\theta) = -\frac{1}{2}\frac{\varepsilon_t\left(\theta\right)}{h_t(\theta)}\dot{h}_{t,\varphi}(\theta). \tag{2.29}$$

Finally,

$$\begin{aligned}
\frac{\partial^2\varepsilon_t\left(\theta\right)}{\partial r_t \partial\varphi} &= \frac{1}{2}h_t(\theta)^{-1/2}\dot{h}_{t,\varphi}(\theta)(h_t(\theta)+\varphi\varepsilon_t^2\left(\theta\right))^{-1} - \\
&\quad -\left(h_t(\theta)+\varphi\varepsilon_t^2\left(\theta\right)\right)^{-2}\left[\dot{h}_{t,\varphi}(\theta)+\varepsilon_t^2(\theta)+2\varphi\varepsilon_t\left(\theta\right)\dot{\varepsilon}_{t,\varphi}(\theta)\right]h_t(\theta)^{1/2} = \\
&= \frac{1}{2}\frac{\sqrt{h_t(\theta)}}{h_t(\theta)+\varphi\varepsilon_t^2(\theta)}\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)} - \frac{\sqrt{h_t(\theta)}}{(h_t(\theta)+\varphi\varepsilon_t^2(\theta))^2}\left[\dot{h}_{t,\varphi}(\theta)+\varepsilon_t^2(\theta)-\varphi\varepsilon_t^2\left(\theta\right)\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)}\right] = \\
&= \frac{\sqrt{h_t(\theta)}}{\left(h_t(\theta)+\varphi\varepsilon_t^2(\theta)\right)^2}\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)}\left[-\frac{1}{2}h_t(\theta)+\frac{3}{2}\varphi\varepsilon_t^2(\theta)\right] - \frac{\sqrt{h_t(\theta)}}{\left(h_t(\theta)+\varphi\varepsilon_t^2(\theta)\right)^2}\varepsilon_t^2(\theta).
\end{aligned}$$

Then

$$\begin{aligned}
\frac{\partial l_t(\theta)}{\partial\varphi} &= -\left(\left[\frac{\partial\varepsilon_t\left(\theta\right)}{\partial r_t}\right]^{-1}\frac{\partial^2\varepsilon_t\left(\theta\right)}{\partial r_t\partial\varphi} - \varepsilon_t\left(\theta\right)\frac{\partial\varepsilon_t\left(\theta\right)}{\partial\varphi}\right) = \\
&= \frac{1}{(h_t(\theta)+\varphi\varepsilon_t^2(\theta))}\frac{\dot{h}_{t,\varphi}(\theta)}{h_t(\theta)}\left(-\frac{1}{2}(\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))(h_t(\theta)+\varphi\varepsilon_t^2(\theta)) - \varepsilon_t^2(\theta)(h_t(\theta)-\varphi\varepsilon_t^2(\theta)\right) = \\
&= \frac{\varepsilon_t^2(\theta)}{(h_t(\theta)+\varphi\varepsilon_t^2(\theta))^2}\left(-\frac{1}{2}b_{t-1}\left(\theta\right)\left(3-\varepsilon_t^2(\theta)\right)+\varphi(\varepsilon_t^4(\theta)-\varepsilon_t^2(\theta))\right) = \\
&= \frac{\varepsilon_t^2(\theta)}{(b_{t-1}(\theta)+2\varphi\varepsilon_t^2(\theta))^2}\left(-\frac{1}{2}b_{t-1}\left(\theta\right)\left(3-\varepsilon_t^2(\theta)\right)+\varphi(\varepsilon_t^4(\theta)-\varepsilon_t^2(\theta))\right).
\end{aligned}$$

I next show that $E\left[\left.\frac{\partial l_t^\star(\theta)}{\partial\lambda}\right|_{\theta=\theta_0}\right] = 0$, where $\lambda = (w,\alpha,\beta)$ and similarly, that $E\left[\left.\frac{\partial l_t^\star(\theta)}{\partial\varphi}\right|_{\theta=\theta_0}\right] = 0$. I start with the score with respect to λ:

$$\frac{\partial l_t(\theta)}{\partial\lambda} = \frac{\dot{b}_{t-1,\lambda}(\theta)}{(b_{t-1}(\theta)+2\varphi\varepsilon_t^2(\theta))^2}\left(-\frac{1}{2}b_{t-1}(\theta)(1-\varepsilon_t^2(\theta))+\varphi\varepsilon_t^4(\theta)+\varphi\varepsilon_t^2(\theta)\right).$$

I calculate the conditional expectation by direct integration against the density of the ε_t, which is just the standard normal density, i.e.:

$$E\left[\left.\frac{\partial l_t^\star(\theta)}{\partial\lambda}\right|_{\theta=\theta_0}\mathcal{F}_{t-1}\right] = \dot{b}^\star_{t-1,\lambda}(\theta) E\left[\frac{-\frac{1}{2}b^\star_{t-1}(\theta_0)\left(1-\varepsilon_t^2\right)+\varphi\left(\varepsilon_t^2+\varepsilon_t^4\right)}{(b^\star_{t-1}(\theta_0)+2\varphi\varepsilon_t^2)^2}\middle|\mathcal{F}_{t-1}\right] =$$

$$= \frac{1}{4\varphi_0^2\sqrt{2\pi}}\dot{b}^\star_{t-1,\lambda}(\theta)\int\limits_{-\infty}^{\infty}\frac{\xi_1 x^4+\xi_2 x^2+\xi_3}{(x^2+\xi_4)^2}e^{-\frac{x^2}{2}}dx, \quad (2.30)$$

where $x=\varepsilon_t(\theta_0)=\varepsilon_t, \xi_1=\varphi_0,\ \xi_2=\varphi_0(1+\frac{b^\star_{t-1}(\theta_0)}{2\varphi}),\ \xi_3=-\frac{1}{2}b^\star_{t-1}(\theta_0)$ and $\xi_4=\frac{b^\star_{t-1}(\theta_0)}{2\varphi_0}$ and note that $\xi_3=\xi_1-\xi_2$ and $\xi_4=-\xi_3/\xi_1$ so I can now apply Lemma 4 to evaluate :

$$\frac{1}{4\varphi^2\sqrt{2\pi}}\dot{b}^\star_{t-1,\lambda}(\theta)\int\limits_{-\infty}^{\infty}\frac{\xi_1 x^4+\xi_2 x^2+\xi_3}{(x^2+\xi_4)^2}e^{-\frac{x^2}{2}}dx = -\frac{1}{4\varphi^2\sqrt{2\pi}}\dot{b}^\star_{t-1,\lambda}(\theta)\left.\frac{\xi_1^2 e^{-x^2/2}x}{\xi_1(x^2-1)+\xi_2}\right|_{-\infty}^{\infty}=0.$$

I now turn to calculating the expectation of the score function at the true parameter vector φ_0. Recall that:

$$\frac{\partial l_t(\theta)}{\partial\varphi} = \frac{\varepsilon_t^2(\theta)}{(h_t(\theta)+\varphi\varepsilon_t^2(\theta))^2}\Big(-\frac{1}{2}b_{t-1}(\theta)(3-\varepsilon_t^2(\theta))+\varphi(\varepsilon_t^4(\theta)-\varepsilon_t^2(\theta))\Big).$$

And similarly to the above case, I calculate the conditional expectation by direct integration against the density of the ε_t, which is just the standard normal density, i.e.:

$$E\left[\left.\frac{\partial l_t^\star(\theta)}{\partial\varphi}\right|_{\theta=\theta_0}\mathcal{F}_{t-1}\right] =$$

$$= E\left[\frac{\varepsilon_t^2}{(h_t^\star(\theta_0)+\varphi_0\varepsilon_t^2)^2}\Big(-\frac{1}{2}b^\star_{t-1}(\theta_0)(3-\varepsilon_t^2)+\varphi_0(\varepsilon_t^4-\varepsilon_t^2)\Big)\middle|\mathcal{F}_{t-1}\right] =$$

$$= \frac{1}{4\varphi_0^2\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\frac{\chi_1 x^6+\chi_2 x^4+\chi_3 x^2}{(x^2+\chi_4)^2}e^{-x^2/2}dx,$$

where $\chi_1=\varphi_0, \chi_2=\frac{1}{2}b^\star_{t-1}(\theta_0)-\varphi_0$ and $\chi_3=-\frac{3}{2}b^\star_{t-1}(\theta_0),\ \chi_4=\frac{b^\star_{t-1}(\theta_0)}{2\varphi_0}$ and notice that it also follows that $\chi_3=-3(\chi_1+\chi_2)$ and $\chi_4=(\chi_1+\chi_2)/\chi_1$ and I therefore can apply Lemma 5 to evaluate:

$$\frac{1}{4\varphi^2\sqrt{2\pi}}\int\limits_{-\infty}^{\infty}\frac{\chi_1 x^6+\chi_2 x^4+\chi_3 x^2}{\left(x^2+\chi_4\right)^2}e^{-x^2/2}dx = -\frac{1}{4\varphi^2\sqrt{2\pi}}\frac{1}{2}\left.\frac{e^{-x^2/2}x^3}{(x^2+\chi_4)}\right|_{-\infty}^{\infty}=0. \qquad \blacksquare$$

Proof of Theorem 4. Proof follows from Assumptions 1-5, Theorem 3 and application of the Central Limit Theorem for martingale difference sequences, e.g. Hall and Heyde (1980) or Pollard (1984). In what follows I apply Theorem 1 in Chapter 8 of Pollard (1984), for which it suffices to verify the following conditions (for properly standardised martingale difference sequence Y_t)

(C1) $\sum_{t=1}^{T} E\left[Y_t^4\right] \to 0,$

(C2) $\sum_{t=1}^{T} E\left[Y_t^2|\mathcal{F}_{t-1}\right] \xrightarrow{p} V_1.$

In my case $Y_t = \frac{1}{\sqrt{T}}\frac{\partial l_t^\star(\theta_0)}{\partial\theta}$. Since the expressions of $\partial l_t^\star(\theta_0)/\partial\theta$ are quite lengthy, I first establish the intermediate result before proving conditions (C1) and (C2) above. In particular, Lemma 4 below states that $\partial l_t^\star(\theta_0)/\partial\theta$ admits moments of any order.

Lemma 6. *Let the model be described by eq.(2.2)-(2.3). Denote by* $\theta = (\lambda', \varphi)'$, *where* $\lambda := (w, \alpha, \beta)'$ *and* $\partial l_t^\star(\theta)/\partial\theta$ *is given by*

$$\frac{\partial l_t^\star(\theta)}{\partial\theta} = \begin{cases} \frac{\dot{b}^\star_{t-1,\lambda}(\theta)}{\left(b^\star_{t-1}(\theta)+2\varphi\varepsilon_t^{2,\star}(\theta)\right)^2}\Big(-\frac{1}{2}b^\star_{t-1}(\theta)(1-\varepsilon_t^{2,\star}(\theta)) + \varphi\varepsilon_t^{2,\star}(\theta) + \varphi\varepsilon_t^{4,\star}(\theta)\Big) & \textit{for } \lambda \\ \frac{\varepsilon_t^{2,\star}(\theta)}{\left(b^\star_{t-1}(\theta)+2\varphi\varepsilon_t^{2,\star}(\theta)\right)^2}\Big(-\frac{1}{2}b^\star_{t-1}(\theta)(3-\varepsilon_t^{2,\star}(\theta)) + \varphi(\varepsilon_t^{4,\star}(\theta) - \varepsilon_t^{2,\star}(\theta))\Big) & \textit{for } \varphi, \end{cases}$$

where $b^\star_{t-1}(\theta) = \alpha + \beta h^\star_{t-1}(\theta) + \gamma r^2_{t-1}$. *The following holds:*

$$\left\|\frac{\partial l_t^\star(\theta_0)}{\partial\theta}\right\|_d < \infty,$$

where $\|X\|_d^d = E|X|^d$ *for* $d > 0$.

Proof of Lemma 6.

I consider separately the derivatives of the score function with respect to each parameter separately. I start with the score function with respect to the parameter vector $\lambda = (w, \alpha, \beta)$. Recall that that the score function w.r.t. λ is given by:

$$\frac{\partial l_t^\star(\theta)}{\partial\lambda} = \frac{\dot{b}^\star_{t-1,\lambda}(\theta)}{\left(b^\star_{t-1}(\theta)+2\varphi\varepsilon_t^{2,\star}(\theta)\right)^2}\Big(-\frac{1}{2}b^\star_{t-1}(\theta)(1-\varepsilon_t^{2,\star}(\theta)) + \varphi\varepsilon_t^{2,\star}(\theta) + \varphi\varepsilon_t^{4,\star}(\theta)\Big)$$

Then taking expectations of the above equation and evaluating it at θ_0 I get:

$$\begin{aligned}
\left\|\frac{\partial l_t^\star(\theta_0)}{\partial\lambda}\right\|_d &= \left\|\frac{\frac{1}{2}b_{t-1}^\star(\theta_0)[\varepsilon_t^2-1]+\varphi_0\varepsilon_t^2+\varphi_0\varepsilon_t^4}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\dot{b}_{t-1,\lambda}^\star(\theta_0)\right\|_d \leq \\
&\leq \frac{1}{2}\left\|\frac{b_{t-1}^\star(\theta_0)[\varepsilon_t^2-1]}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\dot{b}_{t-1,\lambda}^\star(\theta_0)\right\|_d + \left\|\frac{\varphi_0\varepsilon_t^2+\varphi_0\varepsilon_t^4}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\dot{b}_{t-1,\lambda}^\star(\theta_0)\right\|_d \leq \\
&\leq \frac{1}{2}\left\|\dot{b}_{t-1,\lambda}^\star(\theta_0)\frac{1}{b_{t-1}^\star(\theta_0)}\left(\varepsilon_t^2-1\right)\right\|_d + \varphi\left\|\frac{\varepsilon_t^2+\varepsilon_t^4}{b_{t-1}^{2,\star}(\theta_0)}\dot{b}_{t-1,\lambda}^\star(\theta_0)\right\|_d \leq \\
&\leq \frac{1}{2}\|Q_1\|_d\left\{\|(\varepsilon_t^2-1)\|_d+\varphi\|Q_2\|_d\right\},
\end{aligned}$$

where

$$Q_1 = \frac{1}{b_{t-1}^\star(\theta_0)}\dot{b}_{t-1,\lambda}^\star(\theta_0) \quad \text{and} \quad Q_2 = \frac{\varepsilon_t^2+\varepsilon_t^4}{b_{t-1}^\star(\theta_0)}.$$

Note that $\|Q_2\|<\infty$, provided that $E_{\theta_0}\left(\varepsilon_1^2\right)<\infty$ and $E_{\theta_0}\left(\varepsilon_1^4\right)<\infty$. I now show that the term Q_1 is bounded as well. Note that Q_1 involves $\dot{b}_{t-1,\lambda}^\star(\theta_0)$, where $\lambda=(w,\alpha,\beta)'$. Given that the expressions for derivatives of $b_{t-1}^\star(\theta)$ with respect to w, α and β are different, I consider each of them in turn. Using Lemma 2 and the fact that $h_{t-j}^\star(\theta_0)=b_{t-1-j}^\star(\theta_0)+\varphi_0\varepsilon_t^2$ we get:

$$\left\|\frac{\dot{b}_{t-1,w}^\star(\theta_0)}{b_{t-1}^\star(\theta_0)}\right\|_d = \left\|\frac{1+\sum_{k=0}^{\infty}\beta_0^{k+1}\prod_{j=0}^{k}\frac{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}{b_{t-2-j}^\star(\theta_0)+2\varphi_0\varepsilon_{t-1-j}^2}}{\sum_{i=0}^{\infty}\beta^i\left(w_0+\alpha_0r_{t-i-1}^2\right)+\varphi_0\sum_{i=1}^{\infty}\beta_0^i\varepsilon_{t-i}^2}\right\|_d =$$

$$= \left\|\frac{\sum_{k=0}^{\infty}\beta_0^{k}\prod_{j=0}^{k-1}\frac{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}{b_{t-2-j}^\star(\theta_0)+2\varphi_0\varepsilon_{t-1-j}^2}}{\sum_{i=0}^{\infty}\beta_0^i\left(w_0+\alpha_0r_{t-i-1}^2\right)+\varphi_0\sum_{i=1}^{\infty}\beta_0^i\varepsilon_{t-i}^2}\right\|_d \leq \left\|\frac{\sum_{k=0}^{\infty}\beta_0^{k}\prod_{j=0}^{k-1}\frac{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}}{\sum_{i=0}^{\infty}\beta_0^i\left(w_0+\alpha_0r_{t-i-1}^2\right)+\varphi_0\sum_{i=1}^{\infty}\beta_0^i\varepsilon_{t-i}^2}\right\|_d \leq \frac{1}{w_0}<\infty.$$

I now consider the derivative of $b_{t-1}^\star(\theta_0)$ with respect to α.

$$\begin{aligned}
\left\|\frac{\dot{b}_{t-1,\alpha}^\star(\theta_0)}{b_{t-1}^\star(\theta_0)}\right\|_d &= \left\|\frac{\sum_{k=0}^{\infty}\beta_0^{k}r_{t-1-k}^2\prod_{j=0}^{k-1}\frac{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}{b_{t-2-j}^\star(\theta_0)+2\varphi_0\varepsilon_{t-1-j}^2}}{b_{t-1}^\star(\theta_0)}\right\|_d \leq \\
&\leq \left\|\frac{\sum_{k=0}^{\infty}\beta_0^{k}r_{t-1-k}^2\prod_{j=0}^{k-1}\frac{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}{b_{t-2-j}^\star(\theta_0)+\varphi_0\varepsilon_{t-1-j}^2}}{w_0+\alpha_0\beta^k r_{t-1-k}^2}\right\|_d \leq \frac{1}{\alpha_0}\left\|\sum_{k=0}^{\infty}\frac{\alpha_0\beta_0^k r_{t-1-k}^2}{w_0+\alpha_0\beta_0^k r_{t-i-1}^2}\right\|_d \leq \\
&\leq \frac{1}{\alpha_0}\sum_{k=0}^{\infty}\left\|\left\{\frac{\alpha_0\beta_0^k r_{t-1-k}^2}{w_0}\right\}^{s/d}\right\|_d \leq \frac{\alpha_0^{s/d-1}}{w_0^{s/d}}\left\{E_{\theta_0}\left(\alpha_0r_1^2\right)^s\right\}^{1/d}\sum_{k=0}^{\infty}\beta_0^k<\infty,
\end{aligned}$$

where I use the inequality $x/(1+x)\leq x^s$ for all $x>0$. I now consider the derivative of $b_{t-1}^\star(\theta_0)$

with respect to parameter β:

$$\left\| \frac{\dot{b}^{\star}_{t-1,\beta}(\theta_0)}{b^{\star}_{t-1}(\theta_0)} \right\|_d = \left\| \frac{\sum\limits_{k=0}^{\infty} \beta_0^k h^{\star}_{t-1-k}(\theta_0) \prod\limits_{j=0}^{k} \frac{b^{\star}_{t-2-j}(\theta_0)+\varphi_0\varepsilon^2_{t-1-j}}{b^{\star}_{t-2-j}(\theta_0)+2\varphi_0\varepsilon^2_{t-1-j}}}{b^{\star}_{t-1}(\theta_0)} \right\|_d \leq$$

$$\leq \left\| \frac{\sum\limits_{k=0}^{\infty} \beta_0^k h^{\star}_{t-1-k}(\theta_0) \prod\limits_{j=0}^{k-1} \frac{b_{t-2-j}(\theta_0)+\varphi_0\varepsilon^2_{t-1-j}}{b^{\star}_{t-2-j}(\theta_0)+\varphi_0\varepsilon^2_{t-1-j}}}{b^{\star}_{t-1}(\theta_0)} \right\|_d = \left\| \frac{\sum\limits_{k=0}^{\infty} \beta_0^k h^{\star}_{t-1-k}(\theta_0)}{b^{\star}_{t-1}(\theta_0)} \right\|_d =$$

$$= \left\| \frac{\sum\limits_{k=0}^{\infty} \beta_0^k \left(b^{\star}_{t-2-k}(\theta_0) + \varphi_0\varepsilon^2_{t-1-k} \right)}{w_0 + \beta_0^k \left[b^{\star}_{t-2-k}(\theta_0) + \varphi_0\varepsilon^2_{t-1-k} \right]} \right\|_d \leq \left\| \sum_{k=0}^{\infty} \frac{\beta_0^k \left(b^{\star}_{t-2-k}(\theta_0) + \varphi_0\varepsilon^2_{t-1-k} \right)}{w_0 + \beta_0^k \left[b^{\star}_{t-2-k}(\theta_0) + \varphi_0\varepsilon^2_{t-1-k} \right]} \right\|_d \leq$$

$$\leq \sum_{k=0}^{\infty} \left\| \left\{ \frac{\beta_0^k (b^{\star}_{t-2-k}(\theta_0) + \varphi_0\varepsilon^2_{t-1-k})}{w_0} \right\}^{s/d} \right\|_d = \sum_{k=0}^{\infty} \left\| \left\{ \frac{\beta_0^k \sum\limits_{i=0}^{\infty} \left[w_0 + \alpha_0 r^2_{t-2-k-i} + \varphi_0\varepsilon^2_{t-1-k-i} \right]}{w_0} \right\}^{s/d} \right\|_d \leq$$

$$\leq \sum_{k=0}^{\infty} \sum_{i=0}^{\infty} \left\| \left\{ \frac{\beta_0^k \left[w_0 + \alpha_0 r^2_{t-2-k-i} + \varphi_0\varepsilon^2_{t-1-k-i} \right]}{w_0} \right\}^{s/d} \right\|_d$$

$$\leq \sum_{k=0}^{\infty} \left\| \left\{ \frac{\beta_0^{2k} \left[w_0 + \alpha_0 r^2_{t-2-2k} + \varphi_0\varepsilon^2_{t-1-2k}(\theta_0) \right]}{w_0} \right\}^{s/d} \right\|_d \leq \frac{1}{w_0^{s/d}} \left\{ E_{\theta_0} \left(w_0 + \alpha_0 r_1^2 + \varphi_0 \varepsilon_1^2 \right) \right\}^{s/d} \sum_{k=0}^{\infty} \beta_0^{2k} < \infty,$$

where I again used the inequality $x/(1+x) \leq x^s$ for all $x > 0$ and that $E\left(r_1^2\right) < \infty$ and $E\left(\varepsilon_1^2\right) < \infty$.

I now consider $\partial l_t^\star(\theta)/\partial\varphi$, which is given by:

$$\frac{\partial l_t^\star(\theta)}{\partial\varphi} = \frac{\varepsilon_t^{2,\star}(\theta)}{\big(b_{t-1}^\star(\theta)+2\varphi\varepsilon_t^{2,\star}(\theta)\big)^2}\bigg(-\frac{1}{2}b_{t-1}^\star(\theta)(3-\varepsilon_t^{2,\star}(\theta))+\varphi(\varepsilon_t^{4,\star}(\theta)-\varepsilon_t^{2,\star}(\theta))\bigg).$$

Therefore,

$$\begin{aligned}
\left\|\frac{\partial l_t^\star(\theta_0)}{\partial\varphi}\right\|_d &= \left\|\frac{\varepsilon_t^2}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\bigg(-\frac{1}{2}b_{t-1}^\star(\theta_0)(3-\varepsilon_t^2)+\varphi_0(\varepsilon_t^4-\varepsilon_t^2)\bigg)\right\|_d = \\
&= \left\|\frac{1}{2}\frac{\left[b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2\right](\varepsilon_t^2-3)}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}+\frac{2\varphi_0\varepsilon_t^2}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\right\|_d \\
&\leq \frac{1}{2}\left\|\frac{\left[b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2\right](\varepsilon_t^2-3)}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)^2}\right\|_d = \\
&= \frac{1}{2}\left\|\frac{(\varepsilon_t^2-3)}{(b_{t-1}^\star(\theta_0)+2\varphi_0\varepsilon_t^2)}\right\|_d < \infty. \qquad \blacksquare
\end{aligned}$$

Verifying condition (C1). Follows from Lemma 5 for $d=4$.

Verifying condition (C2). To verify condition (C2) it suffices to show the following two conditions:

(i) $\sum_{t=1}^T E\left[Y_t^2\right] \xrightarrow{p} V_1$,

(ii) $\sum_{t=1}^T \Big(E\left[Y_t^2|\mathcal{F}_{t-1}\right]-E\left[Y_t^2\right]\Big) \xrightarrow{p} 0$.

Condition (i) follows by applying Lemma 5 for $d=2$ and applying the LLN for stationary and ergodic sequences. Condition (ii) follows from the fact that

$$E\bigg\{\sum_{t=1}^T\Big(E\left[Y_t^2|\mathcal{F}_{t-1}\right]-E\left[Y_t^2\right]\Big)\bigg\}^2 \to 0,$$

which completes the proof of Theorem 4. $\blacksquare$

I next derive the exact form of the asymptotic variance $\mathbb{V}_{\theta_0}$ in Theorem 4. I start by deriving the second derivative of the (negative) of the log-likelihood $l_t^\star(\theta)$, starting with $\partial l_t^\star(\theta_0)/\partial\varphi$. For simplicity I write $\varepsilon_t^{2,\star}(\theta)$ as simply $\varepsilon_t^2(\theta)$.

$$\frac{\partial l_t^\star(\theta)}{\partial\varphi} = \frac{1}{2}\frac{\varepsilon_t^2(\theta)}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)^2}\left[h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi\left(\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta)\right)\right]. \tag{2.31}$$

I make use of the following notation:

$$A(\theta) := \frac{\varepsilon_t^2(\theta)}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)^2} \quad \text{and} \quad B(\theta) := h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi\left(\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta)\right) \tag{2.32}$$

Using this notation it then follows:

$$\frac{\partial^2 l_t^\star(\theta)}{\partial\varphi^2} = \frac{1}{2}\left[\frac{\partial A(\theta)}{\partial\varphi}B(\theta) + A(\theta)\frac{\partial B(\theta)}{\partial\varphi}\right], \tag{2.33}$$

where using eq.(2.29) and eq.(2.20) it then follows:

$$\frac{B(\theta)}{\partial\varphi} = \dot{h}_{t,\varphi}^\star(\theta)(\varepsilon_t^2(\theta)-3)+2h_t^\star(\theta)\varepsilon_t(\theta)\dot{\varepsilon}_{t,\varphi}(\theta)+\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta)+\varphi\left(4\varepsilon_t^3(\theta)\dot{\varepsilon}_{t,\varphi}(\theta)+2\varepsilon_t(\theta)\dot{\varepsilon}_{t,\varphi}(\theta)=\right.$$
$$= \dot{h}_{t,\varphi}^\star(\theta)(\varepsilon_t^2(\theta)-3) - h_t^\star(\theta)\varepsilon_t^2(\theta)\frac{\dot{h}_{t,\varphi}^\star(\theta)}{h_t^\star(\theta)} + \varepsilon_t^4(\theta)+\varepsilon_t^2(\theta)+$$
$$+\varphi\left(-2\varepsilon_t^4(\theta)\frac{\dot{h}_{t,\varphi}^\star(\theta)}{h_t^\star(\theta)} - \varepsilon_t^2(\theta)\frac{\dot{h}_{t,\varphi}^\star(\theta)}{h_t^\star(\theta)}\right) = -3h_t^\star(\theta)\frac{\dot{h}_{t,\varphi}^\star(\theta)}{h_t^\star(\theta)}-$$
$$-\varphi\frac{\dot{h}_{t,\varphi}^\star(\theta)}{h_t^\star(\theta)}(2\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta))+\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta) = -\frac{\varepsilon_t^2(\theta)}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)}(3h_t^\star(\theta)+2\varphi\varepsilon_t^4(\theta)+\varphi\varepsilon_t^2(\theta))+$$
$$+\varepsilon_t^4(\theta)+\varepsilon_t^2(\theta) = \frac{h_t^\star(\theta)\varepsilon_t^4(\theta)-2h_t^\star(\theta)\varepsilon_t^2(\theta)-\varphi\varepsilon_t^6(\theta)}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)},$$

and

$$\frac{\partial A(\theta)}{\partial\varphi} = -\frac{\varepsilon_t^4(\theta)}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^3} - \frac{4h_t^\star(\theta)\varepsilon_t^4(\theta)}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4} = -\frac{\varepsilon_t^4(\theta)}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4}\left[5h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)\right].$$

Plugging everything back into eq.(2.33) I get:

$$\frac{\partial^2 l_t^\star(\theta)}{\partial\varphi^2} = \frac{1}{2}\left\{-\frac{\varepsilon_t^4(\theta)(5h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4}\left[h_t^\star(\theta)\varepsilon_t^2(\theta)-3h_t^\star(\theta)+\varphi\varepsilon_t^4(\theta)+\varphi\varepsilon_t^2(\theta)\right]+\right.$$
$$\left.+\frac{\varepsilon_t^2(\theta)(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4}\left[h_t^\star(\theta)\varepsilon_t^4(\theta)-2h_t^\star(\theta)\varepsilon_t^2(\theta)-\varphi\varepsilon_t^6(\theta)\right]\right\} = \frac{1}{2}\frac{\varepsilon_t^2(\theta)}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4}\times$$
$$\times\left\{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)\left[h_t^\star(\theta)\varepsilon_t^4(\theta)-2h_t^\star(\theta)\varepsilon_t^2(\theta)-\varphi\varepsilon_t^6(\theta)\right]-\right.$$
$$\left.-(5h_t^\star(\theta)\varepsilon_t^2(\theta)+\varphi\varepsilon_t^4(\theta))\left[h_t^\star(\theta)\varepsilon_t^2(\theta)-3h_t^\star(\theta)+\varphi\varepsilon_t^4(\theta)+\varphi\varepsilon_t^2(\theta)\right]\right\}=$$
$$=\frac{1}{2}\frac{\varepsilon_t^2(\theta)}{(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^4}\left[h_t^{2,\star}(\theta)\varepsilon_t^2(\theta)(13-4\varepsilon_t^2(\theta))-2\varphi h_t^\star(\theta)\varepsilon_t^4(\theta)(2+3\varepsilon_t^2(\theta))-\varphi^2\varepsilon_t^6(\theta)(1+2\varepsilon_t^2(\theta))\right].$$

I now turn to the derivation of $\partial^2 l_t^\star(\theta)/\partial\lambda^2$. Recall that

$$\frac{\partial l_t^\star(\theta)}{\partial\lambda} = \frac{1}{2}\frac{1}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)}\frac{\dot{h}_{t,\lambda}^\star(\theta)}{h_t^\star(\theta)}\left[h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))\right]=$$
$$=\frac{1}{2}\frac{\dot{b}_{t-1,\lambda}^\star(\theta)}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\left[h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))\right].$$

Denote by $C(\theta) = \frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\dot{b}_{t-1,\lambda}^\star(\theta)$ and denote by $D(\theta) = h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+$

$\varepsilon_t^4(\theta))$. Using this notation, I can write:

$$\frac{\partial^2 l_t^\star(\theta)}{\partial\lambda^2} = \frac{1}{2}\left[\frac{\partial C(\theta)}{\partial\lambda}D(\theta) + C(\theta)\frac{\partial D(\theta)}{\partial\lambda}\right], \tag{2.34}$$

where using eq.(2.29) and eq.(2.19) I get:

$$\frac{\partial C(\theta)}{\partial\lambda} = -2(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))^{-3}\left[\dot h_{t,\lambda}^\star(\theta)+2\varphi\varepsilon_t(\theta)\dot\varepsilon_{t,\lambda}^\star(\theta)\right]\dot b_{t-1,\lambda}^\star(\theta)+$$
$$+\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta) = -\frac{2}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^3}\left[\frac{1}{h_t^\star(\theta)}\dot h_{t,\lambda}^\star(\theta)(h_t^\star(\theta)-\varphi\varepsilon_t^2(\theta))\right]\dot b_{t-1,\lambda}^\star(\theta)+$$
$$+\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta) = -\frac{2(h_t^\star(\theta)-\varphi\varepsilon_t^2(\theta))}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^4}\left(\dot b_{t-1,\lambda}^\star(\theta)\right)^2+\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta),$$

and

$$\frac{\partial D(\theta)}{\partial\lambda} = \dot h_{t,\lambda}^\star(\theta)(\varepsilon_t^2(\theta)-1)+2h_t^\star(\theta)\varepsilon_t(\theta)\dot\varepsilon_{t,\lambda}(\theta)+\varphi\left(6\varepsilon_t(\theta)\dot\varepsilon_{t,\lambda}(\theta)+4\varepsilon_t^3(\theta)\dot\varepsilon_{t,\lambda}(\theta)\right) =$$
$$= \dot h_{t,\lambda}^\star(\theta)(\varepsilon_t^2-1)-\varepsilon_t^2\dot h_{t,\lambda}^\star(\theta)+\varphi\left(-3\varepsilon_t^2(\theta)\frac{1}{h_t^\star(\theta)}\dot h_{t,\lambda}^\star(\theta)-2\varepsilon_t^4\frac{1}{h_t^\star(\theta)}\dot h_{t,\lambda}^\star(\theta)\right) =$$
$$= -\frac{1}{h_t(\theta)}\dot h_{t,\lambda}^\star(\theta)\left[h_t^\star(\theta)+\varphi(3\varepsilon_t^2(\theta)+2\varepsilon_t^4(\theta))\right] =$$
$$= -\frac{1}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)}\dot b_{t-1,\lambda}^\star(\theta)\left[h_t^\star(\theta)+\varphi(3\varepsilon_t^2(\theta)+2\varepsilon_t^4(\theta))\right].$$

Finally, plugging everything back into eq.(2.34) it follows that:

$$\frac{\partial^2 l_t^\star(\theta)}{\partial\lambda^2} = \frac{1}{2}\Bigg\{\frac{-2(h_t^\star(\theta)-\varphi\varepsilon_t^2(\theta))[h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))]}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^4}\left(\dot b_{t-1,\lambda}^\star(\theta)\right)^2+$$
$$+\frac{h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta)-\frac{h_t^\star(\theta)+\varphi(3\varepsilon_t^2(\theta)+2\varepsilon_t^4(\theta))}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^3}\left(\dot b_{t-1,\lambda}^\star(\theta)\right)^2\Bigg\} =$$
$$= \frac{1}{2}\frac{h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta)+\frac{1}{2}\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^4}\left(\dot b_{t-1,\lambda}^\star(\theta)\right)^2\times$$
$$\times\Big\{[2(\varphi\varepsilon_t^2(\theta)-h_t^\star(\theta))(h_t^\star(\theta)\varepsilon_t^2(\theta)-h_t^\star(\theta)+3\varphi\varepsilon_t^2(\theta)+\varphi\varepsilon_t^4(\theta))-$$
$$-(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))[h_t^\star(\theta)+3\varphi\varepsilon_t^2(\theta)+2\varphi\varepsilon_t^4(\theta)]\Big\} = \frac{1}{2}\frac{h_t^\star(\theta)(\varepsilon_t^2(\theta)-1)+\varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^2}\ddot b_{t-1,\lambda}^\star(\theta)+$$
$$+\frac{1}{2}\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^4}\left(\dot b_{t-1,\lambda}^\star(\theta)\right)^2\times$$
$$\times\left[h_t^{2,\star}(\theta)(1-2\varepsilon_t^2)-2\varphi h_t^\star(\theta)\varepsilon_t^2(\theta)(6+\varepsilon_t^2(\theta))+\varphi\varepsilon_t^4(\theta)(3\varphi+2\varepsilon_t^2(\theta)-2\varphi\varepsilon_t^2(\theta))\right].$$

I now derive the expression for $\ddot{b}^{\star}_{t-1,\lambda}(\theta)$. From eq.(2.19) I get:

$$\ddot{b}^{\star}_{t-1,\lambda}(\theta) = \left(\dot{h}^{\star}_{t,\lambda}(\theta) - \varphi\varepsilon_t^2(\theta)\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)}\right)\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)} -$$
$$- (h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))\left[\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)}\right]^2 + (h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))\frac{\ddot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)} =$$
$$= -2\varphi\varepsilon_t^2(\theta)\left[\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)}\right]^2 + (h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))\frac{\ddot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)}.$$

Plugging the above expression back into eq.(2.34) I get:

$$\frac{\partial^2 l^{\star}_t(\theta)}{\partial\lambda^2} = -\frac{1}{2}\left\{\frac{h^{2,\star}_t(\theta)(2\varepsilon_t^2(\theta)-1) + \varphi\varepsilon_t^4(\theta)(3\varphi + 4\varphi\varepsilon_t^2(\theta) - 2\varepsilon_t^2(\theta)) + 2\varphi h^{\star}_t(\theta)\varepsilon_t^2(\theta)(5+2\varepsilon_t^2(\theta))}{(h^{\star}_t(\theta)+\varphi\varepsilon_t^2(\theta))^2}\times\right.$$
$$\left.\times\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)} - \frac{h^{\star}_t(\theta)(\varepsilon_t^2(\theta)-1) + \varphi(3\varepsilon_t^2(\theta)+\varepsilon_t^4(\theta))}{h^{\star}_t(\theta)+\varphi\varepsilon_t^2(\theta)}\frac{\ddot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)}\right\}$$

I now turn to the derivation of $\partial^2 l^{\star}_t(\theta)/\partial\varphi\lambda$. I further make use of the following notation:

$$E(\theta) := h^{\star}_t(\theta)\varepsilon_t^4(\theta) - 3h^{\star}_t(\theta)\varepsilon_t^2(\theta) + \varphi(\varepsilon_t^6 + \varepsilon_t^4) \quad \text{and} \quad F(\theta) := (h^{\star}_t(\theta) + \varphi\varepsilon_t(\theta)^2)^2.$$

Then from eq.(2.31) using the notation above it follows that:

$$\frac{\partial l^{\star}_t(\theta)}{\partial\varphi\partial\lambda} = \frac{1}{2}\left[\frac{\partial E(\theta)}{\partial\lambda}\frac{1}{F(\theta)} - \frac{E(\theta)}{F^2(\theta)}\frac{\partial F(\theta)}{\partial\lambda}\right], \tag{2.35}$$

where eq.(2.29) and eq.(2.19) I get:

$$\frac{\partial F(\theta)}{\partial\lambda} = 2(h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))\left[\dot{h}^{\star}_{t,\lambda}(\theta) + 2\varphi\varepsilon_t\dot{\varepsilon}_{t,\lambda}(\theta)\right] =$$
$$= 2(h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))\left[h^{\star}_t(\theta)\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)} - \varphi\varepsilon_t^2(\theta)\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)}\right] =$$
$$= 2(h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))(h^{\star}_t(\theta) - \varphi\varepsilon_t^2(\theta))\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h^{\star}_t(\theta)} = \frac{2(h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta))(h^{\star}_t(\theta) - \varphi\varepsilon_t^2(\theta))}{h^{\star}_t(\theta) + \varphi\varepsilon_t^2(\theta)}\dot{b}^{\star}_{t-1,\lambda}(\theta),$$

and

$$\frac{E(\theta)}{\partial\lambda} = \dot{h}^{\star}_{t,\lambda}(\theta)\varepsilon_t^4(\theta) + 4h^{\star}_t(\theta)\varepsilon_t^3(\theta)\dot{\varepsilon}_{t,\lambda}(\theta) - 3\varepsilon_t^2(\theta)\dot{h}^{\star}_{t,\lambda}(\theta) - 6h^{\star}_t(\theta)\varepsilon_t(\theta)\dot{\varepsilon}_{t,\lambda}(\theta) +$$
$$+ \varphi\left[6\varepsilon_t^5(\theta)\dot{\varepsilon}_{t,\lambda}(\theta) + 4\varepsilon_t^3(\theta)\dot{\varepsilon}_{t,\lambda}(\theta)\right] = -\frac{\dot{h}^{\star}_{t,\lambda}(\theta)}{h_t(\theta)}\varepsilon_t^4(\theta)\left[h^{\star}_t(\theta) + 3\varphi\varepsilon_t^2(\theta) + 2\varphi\right].$$

Plugging all of the above back into eq. (2.35) I get:

$$
\begin{aligned}
\frac{\partial^2 l_t^\star(\theta)}{\partial\varphi\partial\lambda} &= \frac{1}{2}\left[\frac{\partial E(\theta)}{\partial\lambda}\frac{1}{F(\theta)} - \frac{E(\theta)}{F^2(\theta)}\frac{\partial F(\theta)}{\partial\lambda}\right] = \\
&= \frac{1}{2}\Bigg\{ -\frac{\dot{h}^\star_{t,\lambda}(\theta)}{h_t^\star(\theta)}\frac{\varepsilon_t^4(\theta)\left[h_t^\star(\theta)+3\varphi\varepsilon_t^2(\theta)+2\varphi\right]}{h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)^2} - \\
&\quad -\frac{\dot{h}^\star_{t,\lambda}(\theta)}{h_t^\star(\theta)}\frac{2(h_t^\star(\theta)-\varphi\varepsilon_t^2(\theta))\left[h_t^\star(\theta)\varepsilon_t^4(\theta)-3h_t^\star(\theta)\varepsilon_t^2(\theta)+\varphi(\varepsilon_t^6(\theta)+\varepsilon_t^4(\theta))\right]}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^3}\Bigg\} = \\
&= -\frac{1}{2}\frac{\dot{h}^\star_{t,\lambda}(\theta)}{h_t^\star(\theta)}\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^3}\times \\
&\times\Bigg\{\varepsilon_t^4(h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta))(h_t^\star(\theta)+3\varphi\varepsilon_t^2(\theta)+2\varphi)-2(h_t^\star(\theta)-\varphi\varepsilon_t^2(\theta)[h_t^\star(\theta)\varepsilon_t^4(\theta)-3h_t^\star(\theta)\varepsilon_t^2(\theta)+\varphi(\varepsilon_t^6(\theta)+\varepsilon_t^4(\theta))]\Bigg\} = \\
&= -\frac{1}{2}\frac{\dot{h}^\star_{t,\lambda}(\theta)}{h_t^\star(\theta)}\frac{1}{[h_t^\star(\theta)+\varphi\varepsilon_t^2(\theta)]^3}\Bigg\{h_t^{2,\star}(\theta)\varepsilon_t^2(\theta)(6-\varepsilon_t^2)+2\varphi h_t^\star(\theta)\varepsilon_t^4(\theta)(2\varepsilon_t^2-3)+\varphi^2\varepsilon_t^6(4+5\varepsilon_t^2)\Bigg\} \qquad \blacksquare.
\end{aligned}
$$

2.6 Appendix B.

Table 2.1: Parameter estimates for Real-time GARCH(1,1)

Parameter estimates for Real-time GARCH(1,1) with $\varepsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$

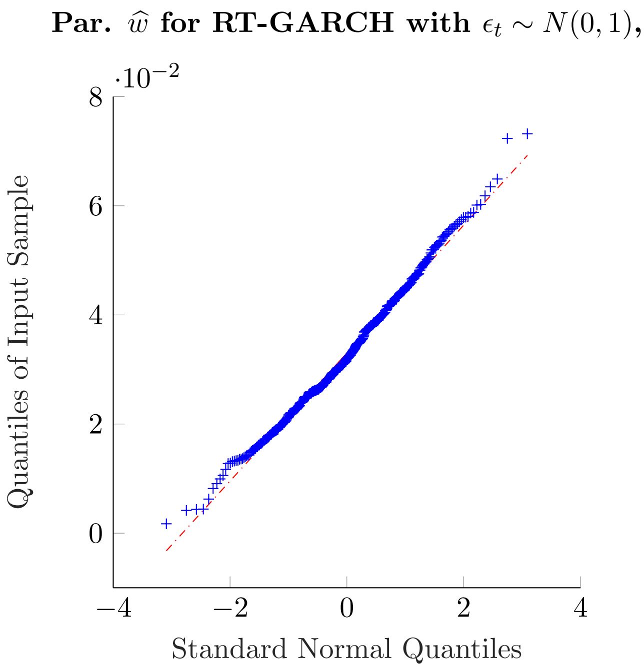
		T = 1 000		**T = 5 000**		**T = 10 000**	
Parameter	True value	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
w	0.0300	0.0280	0.0013	0.0280	0.0020	0.0300	0.0011
β	0.9100	0.9140	0.0115	0.9120	0.0110	0.9105	0.0090
α	0.0300	0.0270	0.0067	0.0335	0.0101	0.0324	0.0075
φ	0.0300	0.0300	0.0014	0.0266	0.0105	0.0268	0.0110

Note: Number of simulations $N = 500$.

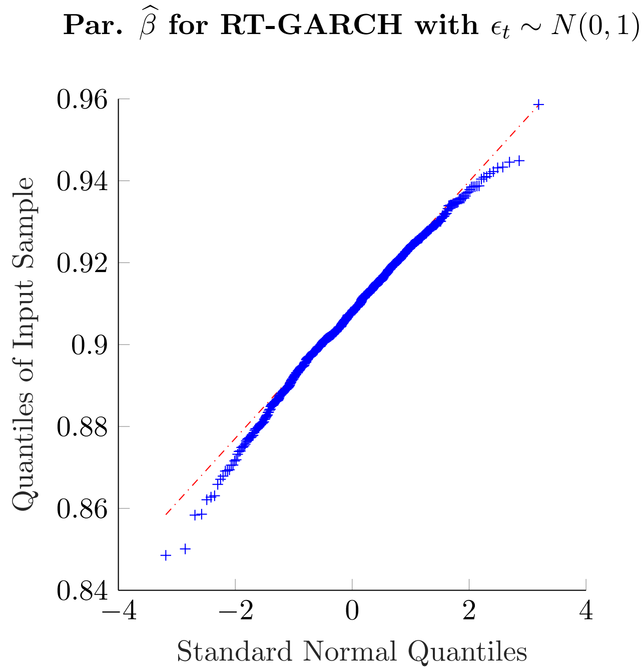
Parameter estimates for Real-time GARCH(1,1) with $\varepsilon_t \sim \mathbf{t_{15}}$

		T = 1 000		**T = 5 000**		**T = 10 000**	
Parameter	True value	Mean	St.dev.	Mean	St.dev.	Mean	St.dev.
w	0.0300	0.0121	0.0479	0.0490	0.0149	0.0370	0.0040
β	0.9100	0.9301	0.0266	0.9272	0.0123	0.9143	0.0085
α	0.03	0.0251	0.0161	0.0273	0.0127	0.0308	0.0061
φ	0.03	0.0776	0.0400	0.0491	0.0327	0.0350	0.0140

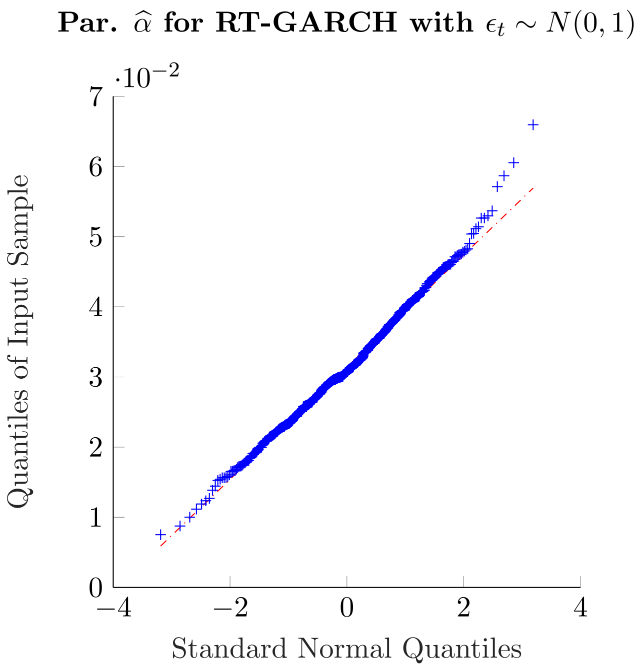
Note: Number of simulations $N = 500$.



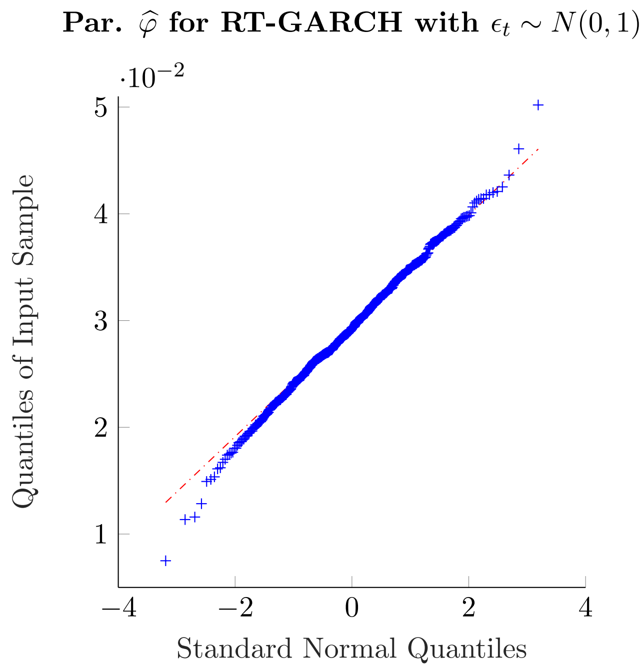
(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{w}$ from Real-time GARCH model when $\varepsilon_t \sim \mathcal{N}(0,1)$.



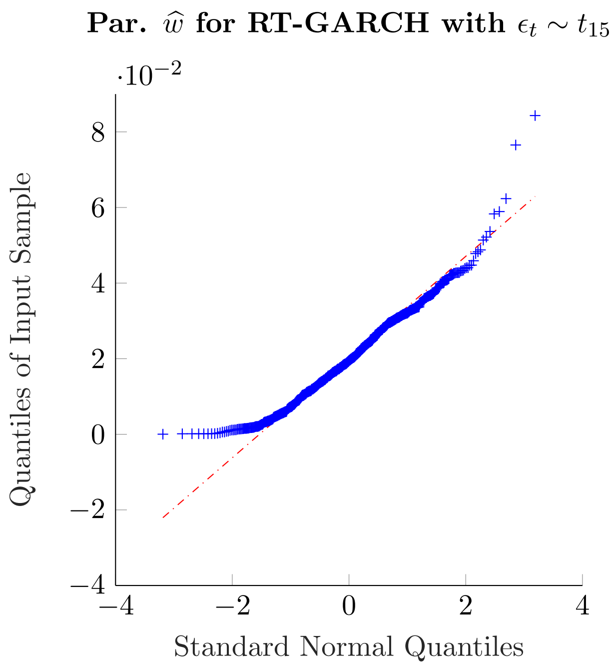
(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\beta}$ from Real-time GARCH model when $\varepsilon_t \sim \mathcal{N}(0,1)$.



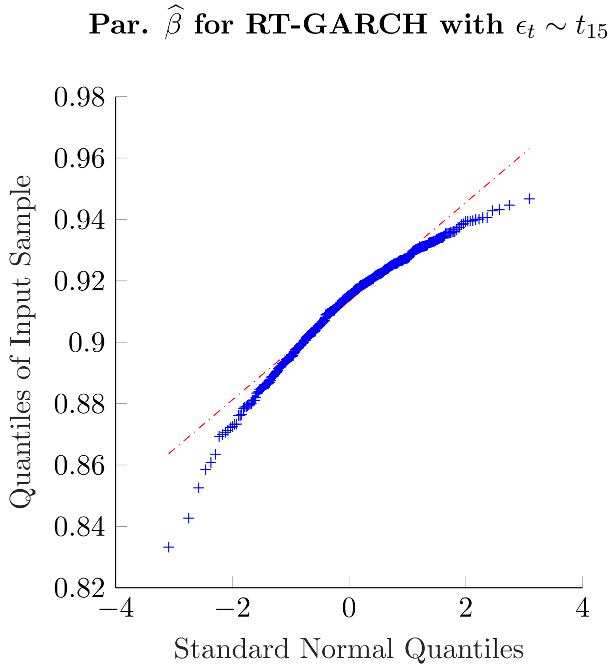
(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{\alpha}$ from Real-time GARCH model when $\varepsilon_t \sim \mathcal{N}(0,1)$.



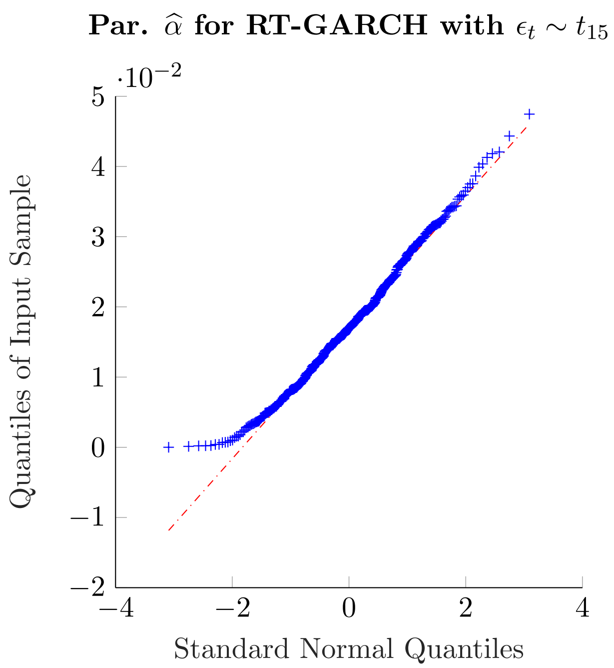
(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\varphi}$ from Real-time GARCH model when $\varepsilon_t \sim \mathcal{N}(0,1)$.



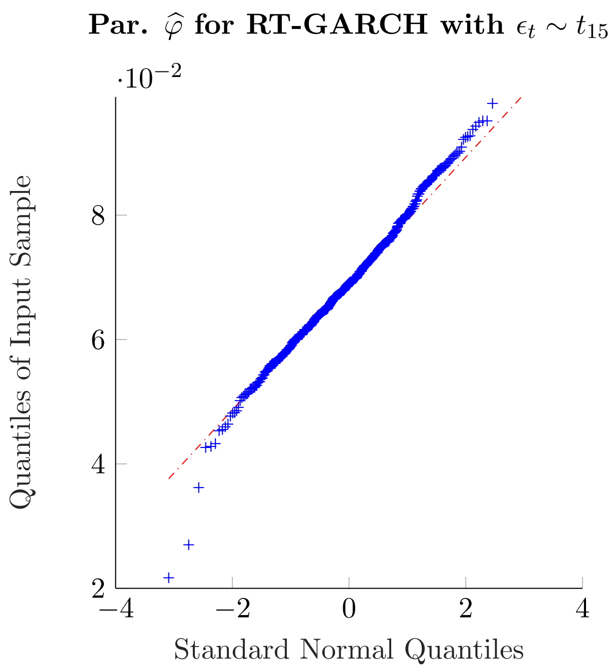
(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{w}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



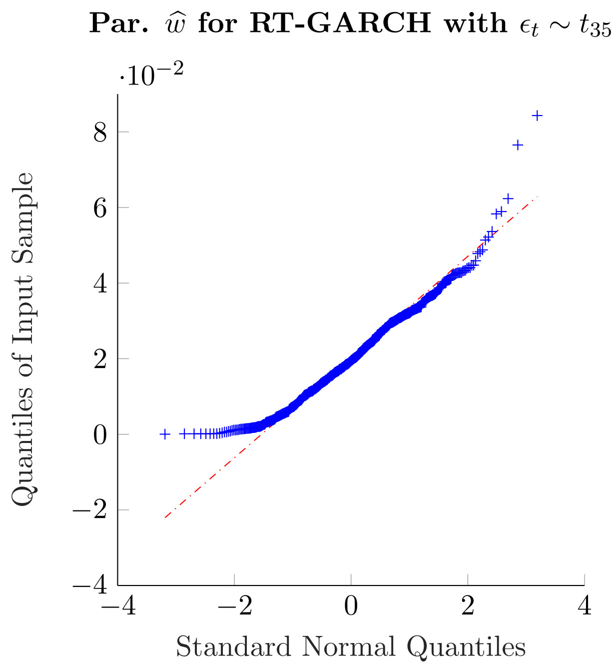
(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\beta}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



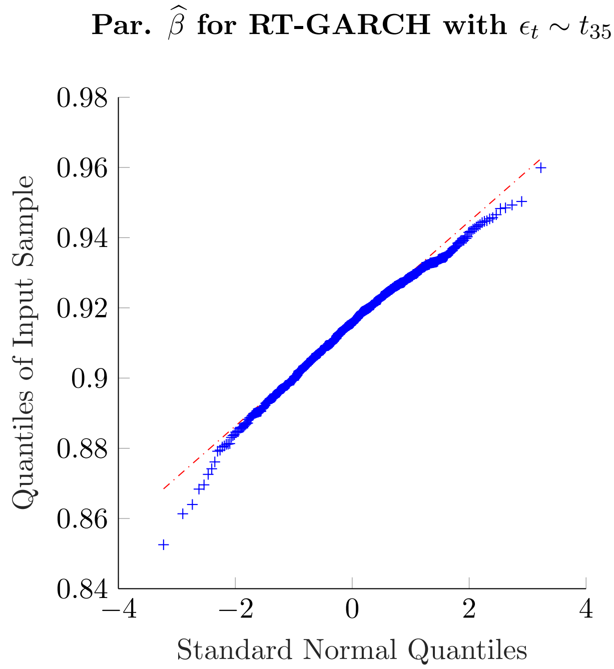
(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{\alpha}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



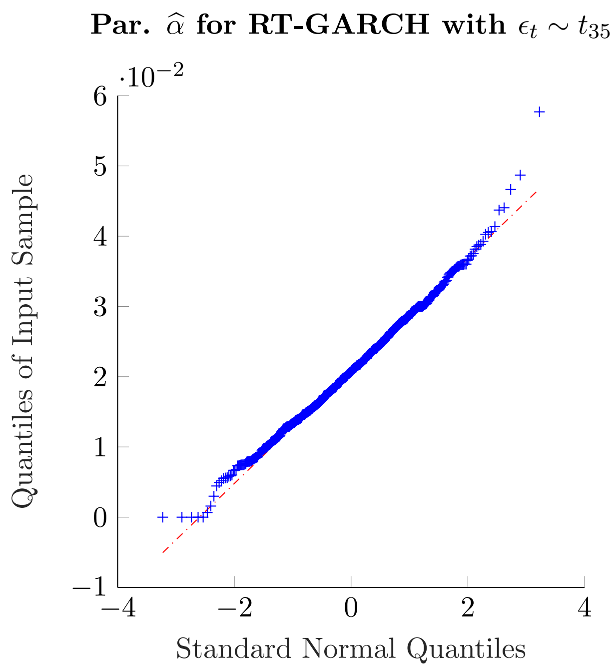
(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\varphi}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



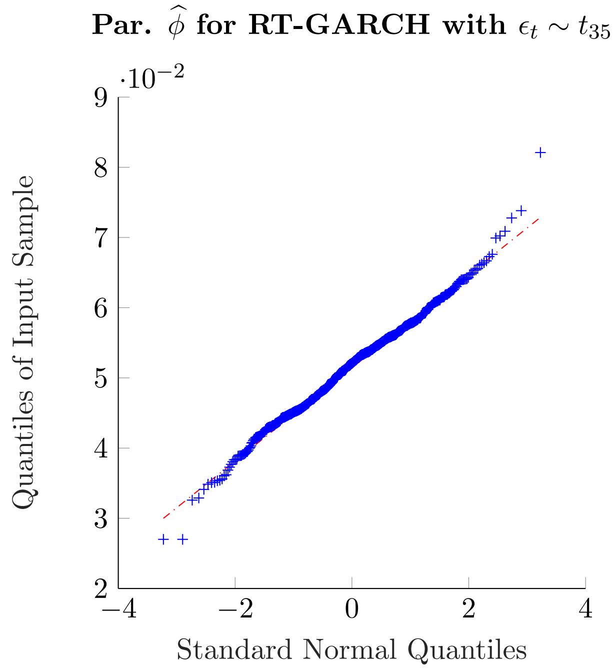
(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{w}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\beta}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



(a) The figure displays the QQ-plot of the QMLE estimator $\widehat{\alpha}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.



(b) The figure displays the QQ-plot of the QMLE estimator $\widehat{\varphi}$ from Real-time GARCH model when $\varepsilon_t \sim t_{15}$.

Chapter 3

Forecast Evaluations in Unstable Environments

3.1 Introduction

In a non-experimental field such as economics, an important way to judge competing models is by comparing their forecasting performance. However, in an unstable environment where relative performance between models can change over time, the existing forecast evaluation methodology can generate spurious and potentially misleading results. An example of this is the well-known "splitting point problem" for out-of-sample forecast evaluation tests. The sample splitting point is used in Diebold-Mariano-type out-of-sample tests to split the sample into the first part, data used for estimation, versus the second part, data used for evaluation. The commonly adopted approach advocates a late sample splitting point, which leaves relatively little data for evaluation and consequently leads to these tests having low power.[1] However beyond this broad guideline, the choice of the splitting point is somewhat arbitrary and left to the discretion of the practitioner. This becomes problematic in an world of changing relative performance. In such a setting, one model may outperform its competition for some *window* of data, but underperform for a different window. Because the splitting point controls the window of data used for evaluation, different splitting points imply different evaluation windows, and the results of these out-of-sample tests may change or completely reverse depending on this arbitrary choice. Consequently, it opens up the possibility of data-mining for practitioners to select favourable splitting points that support their desired hypothesis. Despite these drawbacks, out-of-sample tests are still often preferred to their alternative, in-sample tests. In-sample tests use all available data for both estimation and evaluation, hence they do not suffer a power loss.[2]

[1] See Diebold (2013) for a discussion on this issue and a more recent study by Hirano and Wright (2017) that concludes that current out-of-sample tests perform poorly due to large estimation errors.

[2] Hansen (2008) proposes a methodology for optimal weight selection of forecasts in nested linear models, but it is not clear whether this method can be extended to nonlinear and non-nested models.

However, in-sample tests are prone to spurious results due to over-fitting. Specifically, the ability of a model to fit the data is not necessarily connected to the model's forecasting performance. In fact, Hansen (2010) shows that often, a model's in-sample fit is inversely related to its forecasting performance. See Hansen and Timmermann (2015) for a discussion on the matter.

To demonstrate the two main problems of the existing out-of-sample tests, namely low power and the arbitrary dependence on the splitting point, consider the following real-world example. I forecast the daily variance of IBM returns spanning 2006-2016 using two models: GARCH(1,1) model with Standard normal errors and GARCH(1,1) model with Student-t errors. Each point on the graph below, together with the critical values for the test statistic under the null hypothesis, represents a Diebold-Mariano-type test at that particular splitting point. Variance forecasts are produced via a standard recursive scheme, 5 minute realised volatility calculated from the data is used as a proxy for the "true" variance, and mean squared errors are calculated by averaging squared errors after a particular splitting point. I present the difference in the mean squared errors, ΔMSE_t, and the associated 5% critical values, across a range of splitting point choices, such that the out-of-sample data starts in December 2010, leaving at most 1500 data points for evaluation.

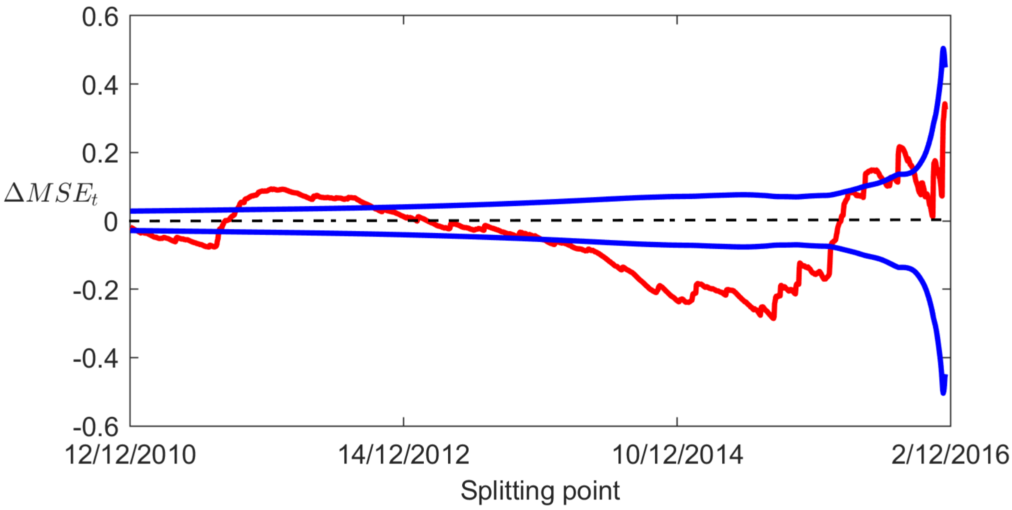


Figure 3.1: The figure displays the *difference* in MSE calculated for GARCH(1,1)-N and GARCH(1,1)-St-t for IBM data, 2006-2016 using recursive forecasting scheme. The MSE for each of the models is taken with respect to 5min RV calculated from the data.

This example is representative of many practical applications. For many plausible choices of the splitting point, the test is not powerful enough to reject in either direction. For other choices of the splitting point, we obtain a rejection in one direction, and for yet other choices, we obtain a rejection in the opposite direction. Hence, depending on the choice of splitting point, all possible conclusions of the test are possible. As the practitioner is often not obliged to show results for all splitting points, in this example it is possible to select any desired outcome.

In this chapter, I propose a new forecast evaluation and selection methodology that is designed explicitly for a world of changing relative performance, where constant relative performance is now a special case. In a changing world, the task of forecast evaluation versus forecast selection do not necessarily overlap. A practitioner may be interested in the question of which model performed better in the past, but it is possible that a different model will outperform for future forecasts. For the purpose of forecast selection, I propose to rank models using two alternative approaches. First, I consider ranking models based on their average past performance. My motivation is that if past performance is indicative of future performance, then a practitioner may want to select models based on average past performance. Importantly, my methodology for evaluating average past performance is robust to the situation of unstable environments, see the discussion in Section 3.2. Second, I consider ranking models based on which model we expect to outperform in the next period. I do this by constructing forecasted probabilities of how likely the forecast loss of one model will be smaller than the forecast loss of another model. A practitioner may then select a model for forecasting next period based on which model is more likely to outperform.

My overall methodology is summarised by a two-step procedure. In the first step, I nonparametrically estimate the time-varying mean and variance for the series of forecast loss differences. In the second step, I utilise these estimates to compare and rank competing models using my two proposed approaches. My statistic measuring average performance aggregates the time-varying means normalised by its time-varying standard deviation, across the entire sample. One therefore can interpret the new test as an aggregated t-test across the whole sample, which is reminiscent to the weighted least squares idea in the standard regression framework. I provide tests for Equal Predictive Ability (EPA) and for Superior Predictive Ability (SPA), which I use to compare and rank models respectively. For my second approach, I construct forecasted probabilities for how likely one model will outperform another based upon the estimates from step one. In addition, I construct *forecast intervals*, which measures the confidence interval of the forecasted probability. In general, my two approaches will often select the same model for forecasting, however this is not always the case. In some applications, a model that performed on average worse over the overall sample may suddenly outperform for a short window towards the end of the sample. In such a situation, my first approach will not select the aforementioned model, however my second approach will. As my second approach is concerned only with the next period performance, the resulting ranking is noisier and subject to change depending on the sample. If a practitioner was interested in selecting a forecast for the next period only, then my second approach should be more relevant, however it is still possible that my forecasted probability is inconclusive while past performance can be accurately compared. In general, I do not advocate one particular approach over the other. Instead I present both approaches and leave the choice to the researcher on which is more relevant for their application. I believe that it shall often be the case that both are insightful, as they address the question of forecast selection from different perspectives, and are both meaningful in their own right.

Related to this work is the paper by Giacomini and White (2006), who develop a conditional version of the unconditional EPA test of Diebold and Mariano (1995). Acknowledging the possible dependence of relative performance on the information set at a given point in time, Giacomini and White (2006) condition the test on a set of covariates, enabling a test for possible variation of relative performance over time. For example, their test rejects their null when models' relative performance depends on a "state of the world" variable, even if the unconditional relative performance is equal. In this case the dependence on the state of the world variable leads to variation in relative performance over time, and in general their test can be thought of as a test for whether we are in a changing world or a constant world (where a rejection of their test is indicative of changing relative performance). Beyond this, and as is acknowledged by the authors, their methodology is not designed for the selection of models for forecasting. Their test informs only whether we reject the conditional null of Equal Predictive Ability, and cannot reliably indicate which model is better in the event of rejection.

Also related to this work, Giacomini and Rossi (2010) were the first to address the problem of what they call "unstable environments", i.e. a world of changing relative performance. They do this by comparing instead the *local* relative performance between two models. This acknowledges directly the possibility for the relative performance of models to change over time, and it is an important step in the literature towards addressing this issue. For the purpose of forecast selection however, the methodology of Giacomini and Rossi (2010) has two shortcomings. First, by focusing their attention on local relative performance, they can only use local data for evaluation, which likely leads to their test having low power. Second, and more importantly, their methodology can only inform the practitioner as to which model was better at a particular point in the past. It is not informative as to which model is better for future forecasts, which is the question of interest to a practitioner who is interested in model selection for future forecasts.[3]

In addition there are the papers by Inoue and Rossi (2012) and by Hansen and Timmermann (2010). They look to address the splitting point problem by bringing to attention the potential data mining of practitioners who search for favourable splitting points. They propose to explicitly mine over all splitting points for the one that is the most favourable for the null hypothesis, and they reevaluate their test statistic at this splitting point with adjusted critical values that account for the bias introduced by mining. However, underlying their test is still the assumption of constant relative performance, and they retain the need for finding the one optimal splitting

[3]This is except in the situation of a clear one-time reversal in relative performance, where one model is clearly better after a sharp structural break. Giacomini and Rossi (2010) consider a version of their test which addresses this scenario. In this special case their methodology can determine which model is superior for future forecasts.

point. In addition, by selecting only the most critical single splitting point they again leverage the result of their test on a particular window of evaluation. As a result, in a world of changing relative performance the mining over splitting points can lead to spurious results. Applied to an example such as the one presented earlier, it may be possible that each model is favoured over the other, where their test selects different splitting points depending on which conclusion they mine for.

The rest of the chapter is organised as follows. In section 3.2 I further discuss the changing relative performance and the two approaches I propose. In section 3.3 I present my theoretical results. Section 3.4 addresses the issue of bandwidth selection for my two-step nonparametric procedure. Section 3.5 describes the bootstrap procedure that is used to approximate the distribution of my new statistics in applications. In section 3.6 I investigate the size and the power of my test under a variety of alternatives as well as the performance of the sign forecasts. I present my applications in section 3.7 and conclude in section 3.8. All proofs of the theoretical results are collected in Appendix B.

Throughout this chapter, the following notation is used. Let $f(x)$ be any function from $\mathbb{R}^d \to \mathbb{R}$, then $\dot{f}(x) = \partial f(x)/\partial x$ and $\ddot{f}(x) = \partial^2 f(x)/\partial x^2$ denote the first and the second derivatives with respect to the argument x respectively. Moreover, $\|g(x)\|_2 = \left(\int |g(x)|^2 dx\right)^{\frac{1}{2}}$ and $\|g(x)\|_2^2 = \int |g(x)|^2 dx$. For a generic non-singular matrix A, A^T denotes its transpose; for a square matrix B, $\operatorname{tr}(B)$ denotes its trace. For any given vector a, $\operatorname{diag}(a)$ creates a diagonal matrix with elements of a along the main diagonal. Finally, $\xrightarrow{p}$ denotes the convergence in probability and $\xrightarrow{d}$ denotes convergence in distribution. All convergences are considered when the sample size $T \to \infty$.

3.2 Model ranking in unstable environments

I start with the framework of two models, although the methodology can be further generalised to many models via pairwise comparisons. Let $\mathcal{A}, \mathcal{B}$ be two models and $\{y_t\}_{t=1}^T$ be the original data. For $i = \mathcal{A}, \mathcal{B}$, let Z_t^i denote any potential predictors used by the forecasting model i and $\widehat{\beta}_t^i$ the parameter estimates of model i at time t, which reflect the model as well as the estimation procedures.[4] I denote the difference in forecast losses at time $t+k$ by $\Delta\mathcal{L}_{t+k}^{\mathcal{AB}} = \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{A}}, Z_t^{\mathcal{A}}\right) - \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{B}}, Z_t^{\mathcal{B}}\right)$, where $\mathcal{L}(\cdot)$ denotes the loss function chosen by the forecaster.[5] In what follows I shall refer to $\Delta\mathcal{L}_{t+k}^{\mathcal{AB}}$ as $\Delta\mathcal{L}_{t+k}$ for simplicity of notation. Note that in general the loss function will be affected by the estimation error. However, given our expanding estimation scheme for constructing the losses, in what follows it is reasonable to

[4] I generically refer to $\widehat{\beta}^{\mathcal{A}}$,$\widehat{\beta}^{\mathcal{B}}$ as parameter estimates, however, depending on whether a parametric, semiparametric or nonparametric model is used $\widehat{\beta}$ will be any estimator used to construct the forecasts.

[5] Throughout the chapter I denote the forecast horizon by k as h will be reserved for denoting the bandwidth.

assume that the estimation error vanishes asymptotically. I also explicitly acknowledge that the mean and variance of $\Delta\mathcal{L}_{t+k}$ might be time-varying. In particular, I define $\mu_{t+k} = \mathbb{E}\left[\Delta\mathcal{L}_{t+k}|\mathbb{X}_t\right]$, where $\mathbb{X}_t$ denotes the set of possible regressors. I make my notation general, so that μ_{t+k} can potentially depend on a set of regressors, in which case μ_{t+k} denotes the conditional mean of the loss difference at time $t+k$. Note that here I refer to $\mathbb{X}_t$ as the set of possible regressors in modelling the conditional mean of the loss differences, and *not* the regressors used to construct forecasts. A natural example of $\mathbb{X}_t$ are the lags of $\Delta\mathcal{L}_t$ as in Giacomini and White (2006). However, it is often the case that we are interested in the unconditional mean which is obtained by setting $\mathbb{X}_t = \varnothing$ for all t. An example of the latter is the commonly applied Diebold-Mariano type tests.

In a world of *constant relative* forecasting performance, i.e. $\mu_{t+k} = \mu$, for all $\mathbb{X}_t$ and all $t \in \{1, \cdots, T\}$, the task of ranking competing models is simple. Specifically, if $\mu < 0$ we say that model $\mathcal{A}$ is better than model $\mathcal{B}$ and vice versa. In such a world, the conclusion of the standard out-of-sample tests does not depend on the evaluation window and hence neither on the choice of the splitting point, although with a too short evaluation window the test shall suffer from low power. Indeed current methodologies explicitly assume constant relative performance, including the tests by West (1996), White (2000), Clark and McCracken (2001, 2005), McCracken (2000, 2007), Hansen (2005), Corradi and Swanson (2007), Hansen et al. (2011), Inoue and Rossi (2012) and by Hansen and Timmermann (2010) and Li and Patton (2017), among others.

However in an unstable environment, i.e. a world of changing relative forecasting performance, the task of model ranking becomes far less obvious. Note that changing relative performance can occur even when the data generating process is stationary (see the example presented in Appendix A). For the example provided in Figure 3.1, the two competing models often overtake each other depending on the evaluation window, and there is no clear dominant choice using any of the previous methodologies. Yet, the question of how to select a model for next period forecasting in such a changing world is of the utmost importance for practitioners.

Giacomini and Rossi (2010) is the first paper in the literature that looks to address the issue of changing relative performance. They propose to rank models at each moment in time by their *local* relative performance. Specifically, for a given forecast horizon k they propose to measure the local mean of losses μ_{t+k} as a sample mean centered around a window of a (fixed) size m, i.e.

$$\widehat{\mu}_{t+k} := \frac{1}{m} \sum_{s=t-m/2}^{t+m/2+1} \Delta\mathcal{L}_{s+k},$$

such that we can test the following null hypothesis:

$$\mathbb{H}_0^{GR}: \quad \mu_{t+k} < 0 \quad \forall\, t \in \{1, \cdots, T-k\} \qquad vs. \qquad \mathbb{H}_1^{GR} \quad \mu_{t+k} \geq 0 \quad \forall\, t \in \{1, \cdots, T-k\}.$$

Giacomini and Rossi's approach is insightful, however as discussed before it still has a few shortcomings if we want to decide with which model we would like to forecast with. We saw previously that each model will outperform the other at some points in time, hence their test shall reject $\mathbb{H}_0^{GR}$ for some t's and accept it for some other t's. Although very useful as an ex-post investigation of the past performance of the competing models, it does not inform the practitioner which model to select for *future* forecasts.

The purpose of this chapter is to provide a methodology to inform model selection in an unstable environment. I do this using two approaches. My first approach is to evaluate past performance, and select models for future forecasts based upon which model performed better in the past. In an unstable environment, the model that outperformed in the past does not necessarily outperform in the future, however it is often the case that past performance is all one can reliably use to compare models. Indeed, the previous forecast evaluation methodology explicitly assumes constant relative performance, and the rationale behind using these tests to select models for forecasting is the same as ours in this first approach. My main contribution in this dimension is to make my evaluation methodology robust to unstable environments, because as we witnessed in the motivating example the previous out-of-sample tests can have various problems in such a situation. My second approach is to forecast the probability that one model shall outperform the other in the next immediate period, i.e. the probability that the sign of the next period loss difference is negative. I choose to forecast the sign of the next period forecast loss difference as opposed to its level as levels can depend on arbitrary factors such as a factor of scaling to the loss function, and it is not clear what kind of a difference in levels constitutes a significant deviation (see Giacomini and White (2006) for a simple application of their framework to level forecasting). Meanwhile, the sign of the loss difference reflects a binary comparison, and indeed the sign for a particular comparison is the same across all symmetric loss functions. This latter approach is conceptually more appropriate to an unstable environment, however in general forecasting this probability is noisier and subject to change depending on the sample. Another limitation of my second approach is that a practitioner may not want to update in every period the model that they choose to forecast with. In which case, they may want to select the model that performed on average better over the entire history, versus the model that is likely to perform better in the immediate next period. As mentioned in the introduction, I believe both approaches to be informative in different situations.

My second approach is more immediate and self evident, however for the first approach there are potentially many ways one could compare past performance. Importantly, we would like the new test to be robust to the various problems of the previous methodology. My first innovation is to use the (near) entire series of forecast losses to construct my statistic, which extends the evaluation to (nearly) the whole sample. This makes my test more powerful, and it makes the result of my test no longer reliant on the arbitrary choice of the sample splitting point. My metric by which I compare past performance is defined as the sum of weighted expected relative forecast losses across the entire sample, where the weighting is given by the time-varying standard deviation of the forecast losses at that point in time. I offer my weighting as my second innovation. My metric to measure past performance belongs to a general class of metrics, which encompasses most of the current methodologies (the general class is also formally defined below). Although my particular metric is just one of many, I argue that it has several attractive features that make it insightful to consider. With my weighting, the forecast losses at the beginning of the sample which come with the largest estimation error are naturally down weighted. Moving towards the end of the sample, more data is used for the estimation leading to lower estimation error, therefore later losses receive a larger weight. My metric is the first to accommodate this explicitly, i.e. that different forecasting losses shall be weighted differently. The motivation for my weighting is to reduce the variance of the statistics, which leads to higher power.

I first define the general class of metrics and its associated ranking by the following definition:

Definition 1. *Let $\mathcal{M}$ be a collection of models under consideration and $\mathcal{M} \times \mathcal{M}$ be the set of all possible model combinations from $\mathcal{M}$ and $\mathcal{A}, \mathcal{B} \in \mathcal{M}$. Let $\Delta \mathcal{L}_{t+k} = \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{A}}\right) - \mathcal{L}\left(y_{t+k}, \widehat{\beta}_t^{\mathcal{B}}\right)$ and $\mu_{t+k} \equiv \mathbb{E}\left[\Delta \mathcal{L}_{t+k} | \mathbb{X}_t\right]$, where $\mathbb{X}_t$ denotes the set of possible regressors. Define the following binary relation on $\mathcal{M}$:*

$$\mathcal{R}_T = \left\{ (\mathcal{A}, \mathcal{B}) \middle| \sum_{t=1}^{T-k} w_{t+k} \mu_{t+k} \leq 0, \mathcal{A}, \mathcal{B} \in \mathcal{M} \right\} \subseteq \mathcal{M} \times \mathcal{M},$$

where $\sum_t w_{t+k} \mu_{t+k}$ is the metric, and $\{w_{t+k}\}_{t=1}^{T-k} \geq 0$ is a set of non-negative weights. We say that model $\mathcal{A}$ is currently superior to model $\mathcal{B}$ iff $(\mathcal{A}, \mathcal{B}) \in \mathcal{R}$.

Remark 1. Note that the above general class of rankings encompasses rankings from the following standard tests (and any variations thereof):

- *Diebold-Mariano (1995) test* if we set $w_{t+k} = 1$ for all $t \geq S$ where S is a splitting point of choice and $\mu_{t+k} = \mu$ for all $t \in T$; and $\mathbb{X}_t = \varnothing$;

- *Giacomini and White (2006) test* if we set $w_{t+k} = 1$ for all $t \geq S$ where S is a splitting point of choice and any $\mathbb{X}_t \in \mathcal{F}_{t-1}$, where $\mathcal{F}_{t-1}$ is the information set available to the forecaster at time $t-1$;

- *Giacomini and Rossi (2010) test* if we set $w_{t+k} = 1$ for all $t \in [S + k - \frac{n}{2} + j, S + k + \frac{n}{2} + j]$, where $j = 0, \cdots, T - n + 1 - S - k$, and n is a fixed size of the rolling window and S is the original splitting point of choice; and $\mathbb{X}_t = \varnothing$.

The metric by which I base my statistic is a special case of the general ranking defined above. The definition for my specific metric is as follows:

Definition 2. *Let $\mathcal{M}$ be a collection of models under consideration and $\mathcal{M} \times \mathcal{M}$ be the set of all possible model combinations from $\mathcal{M}$ and $\mathcal{A}, \mathcal{B} \in \mathcal{M}$. Define the following binary relation on $\mathcal{M}$:*

$$\mathcal{R}_T = \left\{ (\mathcal{A}, \mathcal{B}) \middle| \sum_{t=1}^{T-k} w_{t+k}\mu_{t+k} \leq 0, \mathcal{A}, \mathcal{B} \in \mathcal{M} \right\} \subseteq \mathcal{M} \times \mathcal{M},$$

where $\sum_t w_{t+k}\mu_{t+k}$ is the metric, and $\{w_{t+k} \propto \phi_{t+k}/\sigma_{t+k}\}_{t=1}^{T-k} > 0$ is a set of non-negative weights, where $\sigma_{t+k}^2 = var\left(\Delta\mathcal{L}_{t+k}|\mathbb{X}_t\right)$ and ϕ_{t+k} is a set of (deterministic) given weights. We say that model $\mathcal{A}$ is currently superior to model $\mathcal{B}$ iff $(\mathcal{A}, \mathcal{B}) \in \mathcal{R}$.

Remark 2. The relation in definitions 1 and 2 is transitive, reflexive and antisymmetric, and it generates a partial ordering on the set $\mathcal{M}$.

Remark 3. An example of ϕ_{t+k} is $\phi_{t+k} = \mathbb{1}\left(t + k \in I\right)$, where I could be a period of interest, e.g. recession times.

I now introduce the null hypothesis for my first approach formally. Firstly, I have the following null of Equal Predictive Ability (EPA$_w$):

$$\mathbb{H}_0^{(1)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k}\mu_{t+k} = 0 \qquad vs. \qquad \mathbb{H}_1^{(1)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k}\mu_{t+k} \neq 0, \tag{3.1}$$

and the following null of the Superior Predictive Ability (SPA$_w$):

$$\mathbb{H}_0^{(2)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k}\mu_{t+k} \leq 0 \qquad vs. \qquad \mathbb{H}_1^{(2)}: \quad \sum_{t=\underline{T}+1}^{T} w_{t+k}\mu_{t+k} > 0, \tag{3.2}$$

where under definition 1, w_{t+k} is a set of weights s.t $\{w_{t+k}\} \geq 0$, and under definition 2 w_{t+k} is proportional to $1/\sigma_{t+k}$, and $\underline{T}$ is the point in the sample where I begin my evaluation. I discuss this further in the latter part of this section, see Figure 3.2. Moreover, the notation EPA$_w$ and SPA$_w$ explicitly acknowledges that these are rather a class of null hypothesis, depending on the chosen weighting scheme.

Remark 4. In practice when using the SPA null for model selection, we select model $\mathcal{A}$ as long as we do not reject the above null. If we reject the above SPA null we select model $\mathcal{B}$.

Remark 5. The new test based on the above null hypotheses is also applicable for nested models, i.e. the new test can handle cases when $\mu_{t+k} = \mathbb{E}\left[\Delta\mathcal{L}_{t+k}\right] = \mu = 0$ for all $t = 1, \cdots, T$, so long as there is non-zero variance everywhere. For example, when the practitioner wants to compare two nested models, say for example AR(1) and AR(2), although they might provide the same estimated mean at each point in time with $\mu_{t+k} = \mathbb{E}\left[\Delta\mathcal{L}_{t+k}\right] = 0$ for all $t = 1, \cdots, T$, it is likely the case that $\text{var}\left(\Delta\mathcal{L}_{t+k}\right) \neq 0$ for all $t = 1, \cdots, T$, and therefore my test shall apply. I shall stress nevertheless that there are situations when my test will not be applicable, e.g. in the situation when variance is zero everywhere.

I next provide an intuition for how my test is related to that of Giacomini and White (2006). Recall that Giacomini and White (2006) test the following conditional moment condition: $\mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$, where $\mathcal{F}_t$ is the information set available to forecaster at time t. Provided that $\{\Delta\mathcal{L}_t, \mathcal{F}_t\}$ is a martingale difference sequence, we may test the more lenient in-sample moment condition[6]: $\mathbb{H}_0 : \ \mathbb{E}\left[\Delta\mathcal{L}_{t+1}h_t\right] = 0$, such that $h_t \in \mathcal{F}_t$. The authors recommend to set $h_t = (1, \Delta\mathcal{L}_t)'$. With such a specification, in a regression framework this translates to the following:

$$\Delta\mathcal{L}_t = \alpha + \beta\Delta\mathcal{L}_{t-1} + \varepsilon_t, \qquad \text{and} \qquad \mathbb{H}_0 : \ \alpha = 0 \quad \cap \quad \beta = 0.$$

It is therefore a joint test of the loss functions having the zero mean and absence of serial correlation at first lag. Existence of serial correlation at the first lag would be indicative of changing relative performance, as it is no longer the case that $\mu_t = \mu$ for all t. In general, if we reject their null of $\mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$ due to a dependence of the above moment condition on $\mathcal{F}_t$, this can be considered as evidence of changing relative performance due to a changing information set, $\mathcal{F}_t$. It is still possible that a rejection occurs in a world of constant relative performance, take for example a case where $\mu_t = \mu > 0$ for all t. We possibly may identify changing relative performance as when the Diebold-Mariano (1995) test does not reject, which is indicative of insufficient evidence against zero unconditional mean, however Giacomini and White (2006) test does reject, indicating the rejection is likely due to changing relative performance.

Given my modelling framework, which I shall discuss in the next section, for a particular choice of $\mathbb{X}_t$ one can test a more general null hypothesis that all time-varying coefficients in the regression of $\Delta\mathcal{L}_t$ on $\mathbb{X}_t$ are zero for all $t \in T$. This methodology can be viewed as a way of implementing the conventional Mincer-Zarnowitz (1969) regressions in unstable environments, and it shall nest the Giacomini and White (2006) test as a special case. These tests can also be interpreted as a generic version of the existing forecast rationality tests for a particular choice of $\mathbb{X}_t$. This issue is left for future research.

In what follows, I describe my method for constructing forecast losses, which I do using a standard recursive scheme. Note that contrary to the existing out-of-sample tests, for my metric I need to construct losses for the entire sample, and not just a short evaluation window towards the end of sample.

[6] Meaning that rejection of the null $\mathbb{H}_0 : \ \mathbb{E}\left[\Delta\mathcal{L}_{t+1}h_t\right] = 0$ leads to rejection of $\mathbb{H}_0 : \ \mathbb{E}\left[\Delta\mathcal{L}_{t+1}|\mathcal{F}_t\right] = 0$.

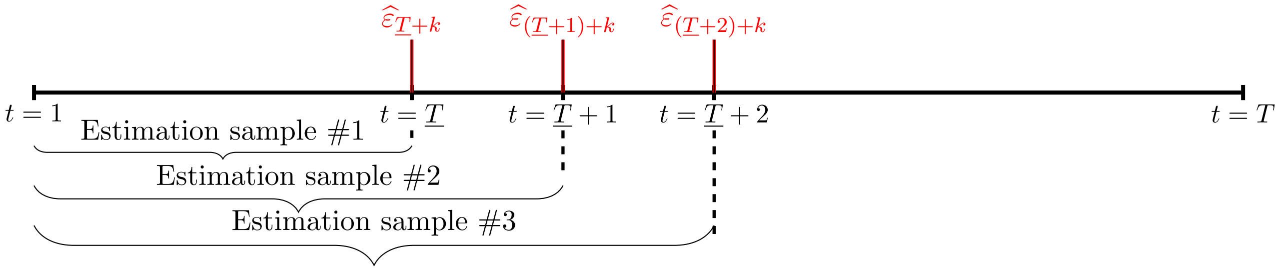


Figure 3.2: Construction of the time series of the forecast errors for a single model.

The pseudo-out-of-sample forecast made at time t for period $t+k$ is compared with the realised value in period $t+k$, which when differenced gives the forecast error of period t. The loss function is then applied to this error which gives the forecast loss of period $t+k$. The recursive scheme calculates the forecast loss using parameter estimates based on all data up until time t. It is recursive because with each new period the model is re-estimated to include the new data. I use all of the forecast losses except for a small initial period of length $\underline{T}$. I do this by always reserving $\underline{T}$ periods for estimation, which can for example be taken to be $\underline{T}=100$.

After the time series of forecast losses is constructed for each model, we may now compute the loss differences for a pair of models, $\mathcal{A}$ and $\mathcal{B}$:

$$\Delta\mathcal{L}_{t+k} = \mathcal{L}\left(\widehat{\varepsilon}^{\mathcal{A}}_{t+k}\right) - \mathcal{L}\left(\widehat{\varepsilon}^{\mathcal{B}}_{t+k}\right), \tag{3.3}$$

where $\mathcal{L}(\cdot)$ is the chosen (by the researcher) loss function. For example, for the conventional squared error loss function it simply becomes

$$\Delta\mathcal{L}_{t+k} = \left(\widehat{\varepsilon}^{\mathcal{A}}_{t+k}\right)^2 - \left(\widehat{\varepsilon}^{\mathcal{B}}_{t+k}\right)^2. \tag{3.4}$$

Following the construction of the loss differences, we may next proceed to the theoretical results, which I discuss in the next section.

3.3 Theoretical results

Once I have constructed the time series of $\Delta\mathcal{L}_t$, I shall from now on only work with this time series and not the original data. For ease of notation I shall say that $\Delta\mathcal{L}_t$ ranges from $t=1,2,\cdots,T$, although the length of this series T is different to the length of the original data y_t. The new T is equal to the original $T-\underline{T}-k+1$.

I model $\Delta\mathcal{L}_t$ as a locally stationary process and allow its mean and variance to change smoothly over time. In particular, I model $\Delta\mathcal{L}_t$ as a function of time and its own lags only. The rationale behind such a modelling framework is as follows. It is a well-established fact that due to estimation error, $\Delta\mathcal{L}_t$ exhibits serial correlation, see e.g. Bollerslev et. al. (2016).

This motivates the autoregressive structure of my model for $\Delta\mathcal{L}_t$. However, in general we do not expect the difference in losses to depend on any other regressors. Therefore, in order to be as agnostic as possible, I do not impose any additional structure on $\Delta\mathcal{L}_t$. As in the literature on locally stationary processes, I make $\Delta\mathcal{L}_t$ depend on the rescaled time points t/T rather than real time t, forming therefore a triangular array, $\{\Delta\mathcal{L}_{t,T} : t = 1, \cdots, T\}$. This rescaling is necessary to justify the properties of the resulting estimation procedures as I will be using the infill asymptotics. So, suppose that we observe the time series of forecast loss differences $\{\Delta\mathcal{L}_{t,T}\}$, $t = 1, 2, \cdots T$. The process $\{\Delta\mathcal{L}_{t,T}\}_{t=1}^{T}$ is assumed to follow an autoregressive model with time-varying coefficients which is given by:

$$\Delta\mathcal{L}_{t,T} = \rho_{t,T}^{0} + \sum_{j=1}^{d} \rho_{t,T}^{j} \Delta\mathcal{L}_{t-j,T} + \xi_{t,T}, \qquad t = 1, \cdots, T, \tag{3.5}$$

where $\mathbb{E}\left[\xi_{t,T} | \mathbb{X}_{t,T}\right] = 0$ with $\mathbb{X}_{t,T} = (1, \Delta\mathcal{L}_{t-1,T}, \Delta\mathcal{L}_{t-2,T}, \cdots, \Delta\mathcal{L}_{t-d,T})^{T}$ and $\rho_{t,T}^{j}$, $j = 1, \cdots, d$ are deterministic functions of time.[7] I use the following rescaling method. Let for each $j \in J$, $\rho^{j}(\cdot)$ be a function on $[0, 1]$ and let

$$\rho_{t,T}^{j} = \rho^{j}\left(t/T\right), \qquad t = 1, \cdots, T.$$

The notation $\rho_{t,T}^{j}$ indicates that $\rho_{t,T}^{j}$ depends on the sample size T and the domain of $\rho^{j}(\cdot)$ becomes more dense in t/T as $T \to \infty$. In other words, the time-varying coefficient functions $\rho^{j}(\cdot)$ does not depend on the real time t but rather on the rescaled time points t/T. Note that model (3.5) is general in the sense that I do not restrict the regressors to be strictly stationary. Instead, I allow the triangular array $\mathbb{X}_{t,T}$ to be locally stationary in the following sense.

Definition 3. *(Vogt, 2012). The process $\{X_{t,T}\}$ is locally stationary if for each rescaled time point $u \in [0, 1]$ there exists an associated process $\{X_t(u)\}$ with the following two properties:*

i) $\{X_t(u)\}$ is strictly stationary;

ii) it holds that

$$\left\| X_{t,T} - X_t(u) \right\| \leq \left(\left| \frac{t}{T} - u \right| + \frac{1}{T} \right) U_{t,T}(u) \qquad a.s.,$$

where $\{U_{t,T}(u)\}$ is a process of positive variables satisfying $\mathbb{E}\left[(U_{t,T}(u))^{\rho}\right] < C$ for some $\rho > 0$ and $C < \infty$ independent of u, t and T. $\|\cdot\|$ denotes an arbitrary norm on $\mathbb{R}^{d}$.

In addition, the error process $\{\xi_{t,T} : t = 1, \cdots, T\}$ is assumed to have the martingale

[7] In the chapter I develop the theory for the general time-varying AR(d) model. However, in all applications I will restrict it to be a simple AR(1) model. I believe it is general, yet simple enough to account for serial correlation of the estimation error in the loss differences $\Delta\mathcal{L}_{t,T}$. I therefore suggest that the reader, unless for having a good reason for an alternative specification, always uses the AR(1) model.

difference property, i.e. for all $t = 1, \cdots, T$

$$\mathbb{E}\left[\xi_{t,T} | \{\mathbb{X}_{s,T} : s \leq t\}, \{\xi_{s,T} : s < t\}\right] = 0 \tag{3.6}$$

Although the above condition rules out autocorrelation in the error terms, it allows for heteroskedasticity. For example (3.6) is satisfied by residuals of the form:

$$\xi_{t,T} = \sigma_{t,T}\varepsilon_t = \sigma\left(\frac{t}{T}\right)\varepsilon_t,$$

where $\sigma(\cdot)$ is a time-varying volatility function and $\{\varepsilon_t\}$ is a martingale difference process with the variance normalized to 1. I impose the martingale difference structure on the regression error terms as, i) this allows to relax the stronger condition of the i.i.d. errors, yet technically convenient as one can use CLT for martingale differences in the proofs, ii) the correlation of $\Delta\mathcal{L}_{t,T}$ is already accounted for by imposing an autoregressive structure. Therefore, I can rewrite my model (3.5) as follows:

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}_{t,T}^T \rho\left(t/T\right) + \sigma\left(t/T\right)\varepsilon_t, \tag{3.7}$$

where $\rho(t/T) = (\rho_0(t/T), \rho_1(t/T), \cdots, \rho_d(t/T))^T$ and $\mathbb{X}_{t,T} = (1, \Delta\mathcal{L}_{t-1,T}, \Delta\mathcal{L}_{t-2,T}, \cdots, \Delta\mathcal{L}_{t-d,T})^T$. Time time-varying coefficient function $\rho(t/T)$ and time-varying volatility function $\sigma(t/T)$ can be estimated by nonparametric kernel techniques. In particular, using the notation $K_h(\cdot) = K(\cdot/h)/h$ for the kernel function, I estimate the model (3.7) in the following way:

Step 1 : First estimate the mean function via the local linear nonparametric estimator. In particular, define the following locally weighted least-squares objective:

$$\widehat{\theta}\left(u\right) = \arg\min_{\theta} \sum_{t=1}^{T} K_{h_1}\left(t/T - u\right)\left(\Delta\mathcal{L}_{t,T} - \mathbb{Z}_{t,T}^T\theta\right)^2,$$

where $\mathbb{Z}_{t,T} = \left(\mathbb{X}_{t,T}, \mathbb{X}_{t,T}\left(t/T - u\right)\right)^T$ and $\theta = \theta(t/T) = (\rho(t/T), \dot{\rho}(t/T))^T$.

Step 2 : Define the estimated error term $\widehat{\xi}_{t,T} = \Delta\mathcal{L}_{t,T} - \mathbb{Z}_{t,T}^T\widehat{\theta}(t/T)$. Then estimate the conditional variance $\sigma^2(t/T)$ by running the local constant nonparametric regression of $\widehat{\xi}_{t,T}^2$ on the rescaled time t/T, i.e.

$$\widehat{\sigma}^2(u) = \arg\min_{a} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\left(\widehat{\xi}_{t,T}^2 - a\right)^2.$$

Remark 6. In the second-step estimation I use the local constant estimator, primarily because contrary to local constant estimator $\widehat{\sigma}(u)$ (with non-negative kernel), the local linear estimator $\widehat{\sigma}(u)$ is not guaranteed to be positive.

Remark 7. Although $\widehat{\sigma}(u)$ is a second-step estimator, Fan and Yao (1998) analysed the asymptotic distribution of such an estimator and showed that its asymptotic distribution is identical to that obtained via one-step estimation based on the true errors ξ_t. This is an important result as this allows one to select the optimal bandwidths h_1, h_2 independently based on the conventional one-step procedures. I discuss the bandwidths selection procedure in section 3.4.

Assumption A1

(i) The function ρ is uniformly bounded below one, i.e. $\sup_{u\in[0,1]}\|\rho(u)\| \leq \overline{\rho} < 1$.

(ii) The function $\sigma(\cdot)$ is bounded from above and from below, i.e. there exist constants $C_\sigma < \infty$ and $c_\sigma > 0$ such that $0 < c_\sigma \leq \sigma(u) \leq C_\sigma < \infty$ for all $u \in [0,1]$.

(iii) The functions ρ and σ are Lipschitz continuous with respect to the rescaled time u.

(iv) The residuals $\{\varepsilon_t\}$ is a martingale difference sequence with respect to the information set $\mathcal{F}_{t-1} = \sigma\left(\Delta\mathcal{L}_s, \varepsilon_s | s \leq t-1\right)$. Moreover, ε_t satisfies $\mathbb{E}\left[|\varepsilon_t|^{4+\delta}\right] < \infty$ for some small $\delta > 0$ and are normalised such that $\mathbb{E}\left[\varepsilon_t^2|\mathcal{F}_{t-1}\right] = 1$.

(v) The error term ε_t has an everywhere positive and continuous density f_ε. The density f_ε is bounded and Lipschitz.

Assumption $A1$ lays out the sufficient conditions for establishing that the process $\{\Delta\mathcal{L}_{t,T}\}$ can be locally approximated by $\Delta\mathcal{L}_t(u)$. Moreover one can also show that for each u the process $\{\Delta\mathcal{L}_t(u), t \in \mathbb{Z}\}$, where $\Delta\mathcal{L}_t(u) = \mathbb{X}_t^T(u)\rho(u) + \sigma(u)\varepsilon_t$ has a strictly stationary solution. Assumption $A1$ corresponds to the assumptions $(M1)-(M3)$, $(\Sigma_1)-(\Sigma_3)$ and $(E1)$ in Vogt (2012) under which he establishes these results for a more general class of models, see Theorems 3.1 and 3.2 in Vogt (2012). However, to establish that the process $\{\Delta\mathcal{L}_{t,T}\}$ is geometrically $\beta-$mixing one extra assumption on the density of the error term ε_t is required.

Assumption A2

(i) The density f_ε fulfills the requirement:

$$\int_{\mathbb{R}} |f_\varepsilon(x) - f_\varepsilon(x+\alpha)|dx \leq C_1|\alpha|,$$

where C_1 is a constant such that $C_1 < \infty$.

Assumption $A2$ corresponds to the assumption $(E3)$ in Vogt (2012), which together with Assumption $A1$ allows one to prove that the process $\Delta\mathcal{L}_{t,T}$ is geometrically β-mixing, see Theorem 3.4 in Vogt (2012) for a proof of this result. Similar assumption can be found in e.g. Orbe et. al. (2005), who establishes the mixing property of the time-varying AR(1) process. Finally, below I introduce the rest of the assumptions that will be necessary to present the estimation theory.

Assumption A3

(i) The functions ρ and σ ate twice continuously differentiable with respect to the rescaled time u and have bounded derivatives.

(ii) The kernel K is a second-order kernel, which is bounded symmetric around zero density function that has a compact support, i.e. $K(v) = 0$ for all $|v| > C_2$ with some $C_2 < \infty$. Moreover K is Lipschitz, i.e. $|K(v) - K(v')| \leq L|v - v'|$ for some $L < \infty$ and all $v, v' \in \mathbb{R}$. In addition, K satisfies $\int K(z)dz = 1$, $\lambda_j = \int z^j K(z)dz$ and $\nu_j = \int z^j K^2(z)dz$.

(iii) The bandwidths h_1 and h_2 satisfy the following conditions: as $T \to \infty$, $h_1 \to 0$, $Th_1 \to \infty$ and $Th_1^5 \to 0$. Similarly it holds that as $T \to \infty$, $h_2 \to 0$, $Th_2 \to \infty$ and $Th_2^5 \to 0$.

Assumption $A3$(i) ensures that the resulting estimators $\widehat{\rho}(\cdot)$ and $\widehat{\sigma}^2(\cdot)$ are well-behaved which will allow me to apply the kernel methods as well as the bootstrap methods later on. Assumptions $A3$(ii)-(iii) are standard assumptions on the kernel function and bandwidths, where I take the kernel K to be the second-order kernel. Note that I work with equally spaced time periods, however this is not strictly necessary. For instance, the theory will hold with slight modifications, which I do not present here, for t_i, $i = 1, \cdots, n$ such that $\{t_i/T,\ i = 1, \cdots, n\}$ is dense on a unit interval.

Before stating the first main results, I need to introduce some further notation. I define the following two matrices:

$$\Omega_{t,T} = \mathbb{E}\left[\mathbb{X}_{t,T}\mathbb{X}_{t,T}^T\right], \qquad \text{and} \qquad H = \begin{bmatrix} I_{d+1} & 0 \\ 0 & h_1 I_{d+1} \end{bmatrix},$$

where I_{d+1} is the identity matrix of dimension $(d+1) \times (d+1)$.

Theorem 1. *Let Assumptions (A1)-(A3) hold. Then for any $u \in (0,1)$ it holds that*

$$\sqrt{Th_1}\left(H\{\widehat{\theta}(u) - \theta(u)\} - h_1^2\mathbb{B}_1(u)\right) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_\theta(u)\right),$$

where

$$\mathbb{B}_1(u) = \frac{1}{2}\begin{pmatrix}\lambda_2\ddot{\rho}(u)\\ 0\end{pmatrix} \qquad \textit{and} \qquad \mathbb{V}_\theta(u) = \begin{pmatrix}\nu_0\sigma^2(u)\Omega^{-1}(u) & 0\\ 0 & \lambda_2^{-2}\nu_2\sigma^2(u)\Omega^{-1}(u)\end{pmatrix}.$$

In step two I estimate the variance $\sigma^2(u)$ by running local constant nonparametric regression of squared residuals

$$\widehat{\xi}^2_{t,T} = \left[\Delta\mathcal{L}_{t,T} - \mathbb{Z}^T_{t,T}\widehat{\theta}(t/T)\right]^2, \qquad t = 1, \cdots, T$$

on the rescaled time t/T. Given that my test statistics based on eq. (3.1) or eq. (3.2) aggregates $\widehat{\mu}_t(u)$ over $u \in [0,1]$ weighted by its standard deviation, it becomes necessary to establish the uniform convergence of $\widehat{\sigma}^2(u)$ over the whole support of u rather than just establishing pointwise consistency of $\sigma^2(u)$. The next theorem states the uniform convergence rate for the second-step estimator $\widehat{\sigma}^2(u)$.

Theorem 2. *Let Assumptions (A1)-(A3) hold. Denote by $I_{h_2} = [C_1h_2, 1 - C_1h_2]$, where $C_1 > 0$ such that $C_1h_2 \to 0$ and $1/C_1 \to 0$. Then*

$$\sup_{u\in I_{h_2}}\left|\widehat{\sigma}^2(u) - \sigma^2(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}} + h_2^2\right),$$

and with probability tending to one it also holds that

$$\sup_{u\in I_{h_2}}|\widehat{\sigma}^2(u)| \le C \le \infty,$$

where $C > 0$.

3.3.1 Test Statistics

Provided with my new definition of ranking, it is then straightforward to state the null and alternative Hypotheses. In particular, consider the SPA$_w$:

$$\mathbb{H}_0: \quad \sum_{t=1}^T w_t\mu_t \le 0 \qquad \text{vs} \qquad \mathbb{H}_1: \quad \sum_{t=1}^T w_t\mu_t > 0. \tag{3.8}$$

In line with the discussion in section 3.2, I choose the weights w_t to be inversely proportional to the standard error of the estimate of μ_t. We might want to make it slightly more general by allowing some extra (given) weighting ϕ_t such that $w_t \sim \phi_t/se_t$, see section 3.2 for the detailed motivation of such a weighting. I form the test statistic corresponding to the above null by replacing the unknown quantities with the respective estimators. I first define the local

t-statistic, which I denote by $\widehat{\tau}(u)$:

$$\widehat{\tau}(u) = \frac{\widehat{\mu}(u)}{\widehat{se}(u)} = \frac{\sqrt{Th_1}\mathbb{X}_t^T(u)\widehat{\rho}(u)}{\widehat{\sigma}(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\widehat{\Omega}^{-1}(u)\mathbb{X}_t(u)}}, \tag{3.9}$$

and then the integrated t-statistic is given by:

$$\mathcal{S}_T = \int_0^1 \widehat{\tau}(u)du,$$

or a slightly extended version with an extra (given) weighting $\phi(u)$:

$$\mathcal{S}'_T = \frac{1}{\sqrt{\Phi}}\int_0^1 \phi(u)\widehat{\tau}(u)du, \qquad \text{with} \qquad \Phi = \int_0^1 \phi^2(u)du.$$

An example of $\phi(u)$ is $\phi(u) = \mathbb{1}\,(u \in I)$, where I could be a period of time that forecaster is interested in, for example, recession times. In what follows I analyse the asymptotic behaviour of $\mathcal{S}_T$ under the null as well as under fixed and local alternatives. The fixed alternative hypothesis[8] is given by

$$\mathbb{H}_1: \qquad \int_0^1 \tau(u)du > 0.$$

In addition, to get a rough idea of the power of the test, I further examine a series of local alternatives, i.e. alternatives that converge to $\mathbb{H}_0$ as the sample size T grows. In particular, I define the sequence of functions $\tau_T(u)$ given by:

$$\tau_T(u) = \tau(u) + c_T \triangle(u),$$

where $c_T \to 0$ as $T \to \infty$, the function $\triangle$ is continuous and the quantity $\int_0^1 \tau(u)du$ satisfies the null hypothesis $\mathbb{H}_0$. Under these local alternatives the process $\Delta\mathcal{L}_{t,T}$ is given by

$$\Delta\mathcal{L}_{t,T} = \mathbb{X}_{t,T}^T\rho\,(t/T) + c_T\triangle\,(t/T)\,\sigma\,(t/T)\sqrt{\nu_0\mathbb{X}_{t,T}^T\Omega^{-1}(u)\mathbb{X}_{t,T}/Th_1} + \xi_{t,T}, \tag{3.10}$$

for $t = 1, \cdots, T$, and therefore under (3.10), I move along the following sequence of local alternatives:

$$\mathbb{H}_{1,T}: \qquad \int_0^1 \tau_T(u)du = c_T\int_0^1 \triangle(u)du.$$

The statistic $\mathcal{S}_T$ under $\mathbb{H}_{1,T}$ gets smaller as the sample size increases and therefore the alternatives $\mathbb{H}_{1,T}$ gets closer and closer to $\mathbb{H}_0$ as $T \to \infty$.

---

[8]By fixed alternative we mean that the value of $\int_0^1 \tau(u)du$ is fixed at a particular value.

THEOREM 3. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$ *under* $\mathbb{H}_0$,

$$\sqrt{T}\left(S_T - h_1^2\mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\mathbb{B}_T = \frac{1}{2}\int_0^1 \frac{\lambda_2\mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}}du. \tag{3.11}$$

The next theorem states the asymptotic distribution of the modified statistic $\mathcal{S}_T'$.

THEOREM 4. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample* $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$ *under* $\mathbb{H}_0$,

$$\sqrt{T}\left(S_T' - h_1^2\mathbb{B}_T'\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\mathbb{B}_T' = \frac{1}{2\sqrt{\Phi}}\int_0^1 \frac{\phi(u)\lambda_2\mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}}du \qquad \text{and} \qquad \Phi = \int_0^1 \phi^2(u)du.$$

I now turn to the theoretical results for the fixed and local alternatives. The next theorem states that the bias-corrected statistic $\mathcal{S}_T$ diverges in probability to infinity under $\mathbb{H}_1$. This allows me to establish consistency of the test against fixed alternatives.[9]

THEOREM 5. *Let Assumptions (A1)-(A3) hold. Then under* $\mathbb{H}_1$,

$$S_T - h_1^2\mathbb{B}_T \xrightarrow{p} \int_0^1 \triangle(u)du > 0,$$

where $\mathbb{B}_T$ *is given by eq.(3.11).*

I next examine the behaviour of $\mathcal{S}_T$ under local alternatives. The theorem 6 below states that the asymptotic power of the test against local alternatives of the form $\tau_T(u) = \tau(u) + c_T\triangle(u)$ with $c_T = 1/\sqrt{T}$ and $\int_0^1 \tau(u)du$ satisfying $\mathbb{H}_0$, is constant for all functions $\triangle$ and is determined by $\int_0^1 \triangle(u)du$.

THEOREM 6. *Let Assumptions (A1)-(A3) hold. Let* $c_T = 1/\sqrt{T}$, *then conditional on the*

[9]This result also holds for the modified statistic $\mathcal{S}_T'$.

sample $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$ *under* $\mathbb{H}_{1,T}$,

$$\sqrt{T}\left(S_T - h_1^2\mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}\left(\int_0^1 \triangle(u)du, 1\right),$$

where $\mathbb{B}_T$ *is given by eq.(3.11).*

3.3.2 Sign Forecasting

I now present the theory for sign forecasting. I could also forecast the level of forecast losses, however for the reasons outlined in section 3.2 I believe that the sign is more informative for model selection. Given that my model (3.5) for $\Delta\mathcal{L}_t$ has an autoregressive structure, we may project which model is likely to forecast better in the next period in the following way. Let $\mathcal{F}_{t,T} = \sigma\left(\Delta\mathcal{L}_{s,T}, \varepsilon_{s,T} | s \leq t\right)$ to be the sigma-algebra generated by the history of $\Delta\mathcal{L}_{t,T}$ and $\varepsilon_{t,T}$, and recall that the model for $\Delta\mathcal{L}_t$ is given by

$$\Delta\mathcal{L}_t = \mathbb{X}_t^T\rho(t/T) + \sigma(t/T)\varepsilon_t, \qquad t = 1, \cdots, T \tag{3.12}$$

where with some abuse of notation due to the meaning of T, x^T denotes the transpose of x and $\mathbb{X}_t = (1, \Delta\mathcal{L}_{t-1}, \cdots, \Delta\mathcal{L}_{t-d})^T$ and ε_t is a m.d.s. At the final point in the sample T we would like to predict the sign of $\Delta\mathcal{L}_{T+1}$, i.e. we would like to know:

$$\begin{aligned} Pr\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{F}_T\right) = Pr\left(\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right) + \sigma\left(\frac{T+1}{T}\right)\varepsilon_{T+1} \leq 0 | \mathcal{F}_T\right) = \\ = Pr\left(\varepsilon_{T+1} \leq \frac{-\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right) = F(\varepsilon). \end{aligned} \tag{3.13}$$

Here, given the autoregressive structure of the difference in losses, $\mathbb{X}_{T+1} \in \mathcal{F}_T$. Strictly speaking, I would need to know $\rho\left(\frac{T+1}{T}\right)$ and $\sigma\left(\frac{T+1}{T}\right)$, however given that both $\rho(\cdot)$ and $\sigma(\cdot)$ are smooth and continuous functions of time, it is reasonable to assume that

$$\rho\left(\frac{T+1}{T}\right) \approx \rho\left(\frac{T}{T}\right) = \rho(1) \quad \text{and} \quad \sigma\left(\frac{T+1}{T}\right) \approx \sigma\left(\frac{T}{T}\right) = \sigma(1).$$

It then holds from (3.13) that

$$Pr\left(\Delta\mathcal{L}_{T+1} \leq 0 | \mathcal{F}_T\right) =: F\left(\varepsilon^\star(1)\right),$$

where $\varepsilon^\star(1) := -\mathbb{X}_{T+1}^T\rho\left(1\right)/\sigma(1)$ is the standardised residual from eq.(3.12) at the last point $u \approx T/T = 1$. Conditional on the sample $\{\Delta\mathcal{L}_t\}_{t=1}^T$, I can estimate the conditional probability as follows:

$$\widehat{Pr}\left(\Delta\mathcal{L}_{T+1}\leq 0\right)=\widehat{F}\left(\widehat{\varepsilon}^{\star}(1)\right)=\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\left(\widehat{\varepsilon}_{t}\leq\frac{-\mathbb{X}_{T+1}^{T}\widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right),$$

where $\widehat{\varepsilon}^{\star}(1)$ is an estimate of $\varepsilon^{\star}(1)$. Therefore for a given sample $\{\Delta\mathcal{L}_t\}_{t=1}^{T}$ the practitioner can calculate the probability of $\Delta\mathcal{L}_{T+1}$ of being negative. I state the theoretical result in Theorem 7 below, which allows the practitioner to calculate the probability as well as the confidence intervals for this probability, which I call **forecast intervals**.

THEOREM 7. *Let Assumptions (A1)-(A3) hold. Let $\mathcal{F}_{t,T}=\sigma\left(\Delta\mathcal{L}_{s,T},\varepsilon_{s,T}|s\leq t\right)$ to be the sigma-algebra generated by the history of $\Delta\mathcal{L}_{t,T}$ and $\varepsilon_{t,T}$. Then the forecast of the sign of $\Delta\mathcal{L}_{T+1}$ made at time T is given by*

$$\sqrt{T}\left[\widehat{F}(\widehat{\varepsilon}^{\star}(1))-F(\varepsilon^{\star}(1))-\mathbb{B}_3(1)\right]\xrightarrow{d}\mathcal{N}\Bigg(0,F\left(\varepsilon^{\star}(1)\right)\left(1-F\left(\varepsilon^{\star}(1)\right)\right)\Bigg),\tag{3.14}$$

where

$$\mathbb{B}_3(1)=\frac{f\left(\varepsilon^{\star}(1)\right)}{2\sigma^2(1)}\mathbb{X}_{T+1}(1)\left\{h_1^2\lambda_2\ddot{\rho}(1)\sigma(1)+h_2^2\lambda_2\ddot{\sigma}(1)\right\},$$

and

$$F(\varepsilon^{\star}(1))=Pr(\Delta\mathcal{L}_{T+1}\leq 0)\qquad and\qquad \varepsilon^{\star}(1):=\frac{-\mathbb{X}_{T+1}^{T}\rho\left(1\right)}{\widehat{\sigma}(1)}.$$

Remark 8. The bias term $\mathbb{B}_3(1)$ is due to the estimation error of $\widehat{\varepsilon}(1)$, which itself involves the estimates of $\widehat{\rho}(1)$ and $\widehat{\sigma}(1)$, leading to a particular form of the bias in Theorem 7.

In the simulations (see Figure 3.8) we see that the sign forecasts perform quite well, forecasting near to the true probability. In particular, the sign forecasts improve as we go later in the sample. Because the bandwidth for the first estimation step for ρ is quite small in this particular example, this improvement is not due to estimating ρ more precisely; instead it is due to approximating the c.d.f. of ε better as we go later in the sample, using more data. In general, it looks as if the difficulty of approximating the c.d.f. of ε is greater than the issues surrounding estimating ρ imperfectly. Also, because I am not interested in forecasting the level of the forecast loss difference next period, but rather its sign, my results are somewhat less sensitive to the imprecision caused by using a two-sided kernel. In general, if the p.d.f. at the particular ε^* threshold is small, the probability will not respond much to inaccuracies in ρ. One way to improve forecasts even further would be to use the derivatives of $\widehat{\rho}(1)$ and $\widehat{\sigma}(1)$.

In practice, we are only concerned about making predictions at the last point in time T. However, in one of my simulations and in all of my applications I will be producing pseudo out-of-sample sign forecasting to assess the quality of my procedure. In my simulation, I derive the true probabilities explicitly and compare it with my forecasted probabilities. In my applications where the true probabilities are unknown, I use the following criterion to assess the quality of my forecasts:

$$\widehat{C} := \frac{1}{T - \underline{T}} \sum_{t=\underline{T}}^{T} \left[\mathbb{1} \left(\Delta \mathcal{L}_{t+1} \leq 0 \right) - \widehat{Pr}^{bc} \left(\Delta \mathcal{L}_{t+1} \leq 0 | \Delta \mathcal{L}_t \right) \right], \tag{3.15}$$

where $\underline{T}$ is where I begin my evaluation near the beginning of the sample, e.g. $\underline{T} = 100$, and $\widehat{Pr}^{bc}(\cdot)$ denotes the bias-corrected probability. If the forecasted probabilities were correct, then the criterion above should on average equal to zero. The bias as well as the forecast intervals can be obtained via bootstrap which I discuss in detail in section 3.5.

3.4 Bandwidth selection

In this section I briefly describe how I choose the optimal bandwidths h_1 and h_2. I start with the optimal selection of the first stage estimation bandwidth h_1. The conventional way to choose the optimal bandwidth is to construct the asymptotic mean squared error given by:

$$\text{AMSE}(h_1) = \frac{h_1^4}{4} \mu_2^2 \|\ddot{\theta}(u)\|_2^2 + \frac{\text{tr}\left(\mathbb{V}_\theta(u)\right)}{T h_1},$$

where $V_\theta(u)$ is given in Theorem 1. Then minimising AMSE(h_1) with respect to h_1 provides the optimal bandwidth h_1^{opt} given by:

$$h_1^{opt} = \left\{ \text{tr}\left(\mathbb{V}_\theta(u)\right) \mu_2^{-2} \|\ddot{\theta}(u)\|_2^{-2} \right\}^{-1/5} T^{-1/5} \tag{3.16}$$

However, note that (3.16) involves the unknown quantity $\ddot{\theta}(u)$ that therefore has to be estimated first before the optimal bandwidth can be computed. Several other methods has been proposed in the literature, one of which is multi-fold cross-validation see e.g. Cai, Fan and Yao (2000), Cai, Fan and Li (2000) which takes into account the time-series structure of the data. More precisely, I first partition the data into Q groups (usually $Q = 20$), with the jth group consisting of the data points with indices:

$$d_j = \{Qk + j, \quad k = 1, 2, 3, \cdots\}, \quad j = 0, 1, 2, \cdots, Q - 1.$$

I then fit the model and obtain the estimate of $\widehat{\theta}^{-j}$ using the remaining data after deleting the jth group. Now denote by Y_{-d_j} the fitted values of Y_t using the data with the jth group deleted. Then the cross-validation criterion has the following form:

$$\text{CV}(h_1) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left[Y_i - \widehat{Y}_{-d_j} \right]^2.$$

Alternatively, one can form variants of the cross-validation criteria based on the Pearson's residuals:

$$\mathrm{CV1}(h_1) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left[Y_i \log \left\{ \frac{Y_i}{\widehat{Y}_{-d_j}} \right\} - \{Y_i - \widehat{Y}_{-d_j}\} \right],$$

where in the above one would need to set $0 \log 0 = 0$ to account for the cases when $Y_i = 0$. Finally, another cross-validation criterion can be

$$\mathrm{CV2}(h_1) = \sum_{j=0}^{Q-1} \sum_{i \in d_j} \left[\frac{Y_i - \widehat{Y}_{-d_j}}{\sqrt{\widehat{Y}_{-d_j}}} \right].$$

Minimizing the $\mathrm{CV}(h_1)$ with respect to h_1 then yields the optimal bandwidth h_1^{opt}. In practice, and in general, as established by Cai, Fan and Li (2000) the cross-validation is not too sensitive to the way the data is partitioned. The second-stage estimation procedure of estimating the conditional variance $\sigma^2(u)$ via the local constant estimator is very standard, and the optimal bandwidth h_2^{opt} is estimated via conventional least-squares cross-validation, see e.g. Li and Racine (2007) for details.

3.5 Bootstrapping $\mathcal{S}_T$

Theorems 3-6 allow one to conduct inference for $\mathcal{S}_T$, as the distribution of the test statistics is the simple standard normal distribution. Note also, that the test statistic $\mathcal{S}_T$ is a nonparametric statistic, however through aggregation it converges to $\mathcal{N}(0,1)$ with the standard parametric $\sqrt{T}$ rate. The bias term in Theorems 3-6, however, contains unknown quantities, such as $\ddot{\rho}(u)$. Although it is possible to estimate these unknown quantities, replacing them with the consistent estimates will further result in approximation errors. I therefore choose to bootstrap the statistics, which will automatically allow me to estimate the bias without estimating the unknown quantities. In what follows, I discuss the bootstrap procedure in the context of Theorems 3-6, however the same methodology will be applied to obtain the bias and the forecast intervals in Theorem 7. I set up the fixed regressor wild bootstrap procedure to account for the time series structure of the data. In particular, the wild bootstrap sample, which I denote by $\{\Delta\mathcal{L}^{\star}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^{T}$, where

$$\Delta\mathcal{L}^{\star}_{t,T} = \mathbb{X}^{T}_{t,T}\widetilde{\rho}\,(t/T) + \xi^{\star}_{t,T}, \tag{3.17}$$

and the bootstrap residuals are constructed as follows:

$$\xi^{\star}_{t,T} = \widehat{\xi}_{t,T}\eta_t,$$

where $\widehat{\xi}_{t,T} := \Delta\mathcal{L}_{t,T} - \widehat{\mu}_t(t/T) = \Delta\mathcal{L}_{t,T} - \mathbb{X}_{t,T}^T\widehat{\rho}(t/T)$ are the estimated residuals and $\{\eta_t\}_{t=1}^T$ is a sequence of i.i.d. variables normalized such that it has zero mean and unit variance. I further choose η_t to have a Rademacher distribution. Finally $\widetilde{\rho}_g(\cdot)$ is given by:

$$\widetilde{\rho}_g(u) := \widehat{\rho}(u) - \bar{\rho}, \qquad \bar{\rho} = \int_0^1 \widehat{\rho}(u)du, \tag{3.18}$$

The intuition behind construction of the mean function of $\Delta\mathcal{L}_{t,T}^\star$ given by (3.17)-(3.18) is such that the bootstrapped sample $\{\Delta\mathcal{L}_{t,T}^\star, \mathbb{X}_{t,T}\}_{t=1}^T$ imitates the model under the null hypothesis whether the alternative hypothesis is true or not. Therefore the distribution of the bootstrapped statistic $S_T^\star$, stated below, mimics the distribution of S_T under the null hypothesis regardless whether the null holds or not. Given the bootstrap sample $\{\Delta\mathcal{L}_{t,T}^\star, \mathbb{X}_{t,T}\}_{t=1}^T$, I define the bootstrapped test statistics $\mathcal{S}_T^\star$

$$\mathcal{S}_T^\star = \int_0^1 \frac{\widehat{\mu}_t^\star(u)}{\widehat{se}_t^\star(u)}du, \qquad \widehat{se}_t^\star(u) = \widehat{\sigma}^\star(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\widehat{\Omega}^{-1}(u)\mathbb{X}_t(u)/Th_1},$$

where

$$\widehat{\mu}_t^\star(u) = \widehat{\rho}^\star(u)\mathbb{X}_t^T(u), \quad \text{where} \quad \widehat{\rho}^\star(u) := \arg\min_a \sum_{t=1}^T K_{h_1}(t/T - u)\left(\Delta\mathcal{L}_{t,T}^\star - \mathbb{X}_{t,T}^T a\right)^2,$$

and defining $\widehat{\xi}_{t,T}^\star = \Delta\mathcal{L}_{t,T}^\star - \widehat{\mu}_t^\star$, I further have that $\widehat{\sigma}^\star(u)$ is given by

$$\widehat{\sigma}^\star(u) = \frac{\sum_{t=1}^T K_{h_2}(t/T - u)\left(\widehat{\xi}_{t,T}^\star\right)^2}{\sum_{t=1}^T K_{h_2}(t/T - u)}.$$

The next theorem states that the wild bootstrap described above is consistent.

Theorem 8. *Let Assumptions (A1)-(A3) hold. Then conditional on the sample $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T$ with probability tending to one*

$$\sqrt{T}\left(\mathcal{S}_T^\star - \mathcal{S}^\star - h_1^2\mathbb{B}_T\right) \xrightarrow{d} \mathcal{N}(0,\ 1),$$

where $\mathbb{B}_T$ is given by eq.(3.11). In other words, $\mathbb{P}^\star\left(\mathcal{S}_T^\star - \mathcal{S}^\star - h_1^2\mathbb{B}_T \le x\right) \xrightarrow{p} \Phi(x)$, where $\Phi(x)$ is a Gaussian distribution function with zero mean and variance 1.

Once the wild bootstrap is set up, the size and the power of the test in the next section will then be calculated as follows. I denote by $\mathcal{S}_{T,n}$ the value of the test statistic $\mathcal{S}_T$ in the n-th simulation, and let $\mathcal{S}_{T,n,b}^\star$ be the value of the bootstrap statistics $\mathcal{S}_T^\star$ in the b-th bootstrap sample generated in the n-th simulation. I denote by $G_n^\star$ the empirical distribution function calculated from the sample of the bootstrap values in n-th simulation, i.e. of $\{\mathcal{S}_{T,n,b}^\star\}_{b=1}^B$. Then the actual size of the test statistics can be calculated as follows. Given a fixed nominal size α, for each simulated sample $n \in N$, calculate the $(1-\alpha)$-quantile of $G_n^\star$, denoted by $q_{\alpha,n}^\star$. Finally I compute the actual size and power corresponding to the nominal level α as

$$\frac{1}{N}\sum_{n=1}^{N} \mathbb{1}\left(\mathcal{S}_{T,n} > q_{\alpha,n}^\star\right).$$

3.6 Simulations

In this section I provide the simulation results for the size and power of my test statistic $\mathcal{S}_T$ and demonstrate the sign forecasting methodology. I start by investigating the size of my test $\mathcal{S}_T$.

3.6.1 Test Statistics: Size

For all simulations I set the number of simulations $N = 1000$ and I vary the number of bootstrap replications, B, between $B = 500$, $B = 750$, and $B = 1000$. I start with replicating two alternatives from Giacomini and White (2006) that constitute my null hypothesis. In particular I simulate the loss difference $\Delta\mathcal{L}_t$ as the following AR(1) process:

$$\mathbb{H}_0^{(1)}: \qquad \Delta\mathcal{L}_t = \mu(1-\rho) + \rho\Delta\mathcal{L}_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1) \tag{3.19}$$

For each of $n \in N$ simulations I generate a sequence of loss differences $\Delta\mathcal{L}_t$ of length $T = 150$ according to (3.19), starting from the initial value of $\Delta\mathcal{L}_t$ that equals the difference of squared errors for forecasts of the second log difference of the monthly U.S. consumer price index (CPI), $\mathrm{CPI}_{2016:12}$ implied by two models: i) a white noise; and ii) an AR(1) model for CPI estimated over a window of size $m = 150$ using the data up to 2016:11. Moreover, I consider the scenario with zero unconditional mean and $\rho(0, 0.05, \cdots, 0.9)$.[10] Tables 3.1-3.2 show the simulated actual size for different levels of the nominal size $\alpha = 0.01,\ 0.05,\ 0.10,\ 0.15$.

Table 3.1: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(1)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.072	0.117	0.162
$B = 750$	0.009	0.062	0.107	0.156
$B = 1000$	0.009	0.057	0.104	0.151

[10] I present the results for $\rho = 0.2$ only as varying ρ virtually leaves the results unchanged.

Table 3.2: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(1)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.056	0.101	0.160
$B = 750$	0.011	0.050	0.099	0.162
$B = 1000$	0.011	0.051	0.097	0.155

For the second null hypothesis, also borrowed from Giacomini and White (2006), for $T = 150$ I generate the sequence of loss differences as follows:

$$\mathbb{H}_0^{(2)}: \qquad \Delta\mathcal{L}_t = \frac{\mu}{p(1-p)}(S_t - p) + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.\mathcal{N}(0,1), \tag{3.20}$$

where $S_t = 1$ with probability p and $S_t = 0$ with probability $1-p$, with $p = 0.5$. I thus have that the unconditional mean $\mathbb{E}[\Delta\mathcal{L}_t] = 0$, however

$$\mathbb{E}[\Delta\mathcal{L}_t | S_t] = \begin{cases} \mu/p & \text{if} \quad S_t = 1 \\ -\mu/(1-p) & \text{if} \quad S_t = 0. \end{cases}$$

Tables 3.3-3.4 show the simulated actual size for different levels of the nominal size $\alpha = 0.01,\ 0.05,\ 0.10,\ 0.15$.

Table 3.3: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(2)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.020	0.048	0.108	0.160
$B = 750$	0.018	0.052	0.107	0.154
$B = 1000$	0.015	0.050	0.103	0.150

Table 3.4: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(2)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.012	0.058	0.104	0.150
$B = 750$	0.011	0.050	0.099	0.148
$B = 1000$	0.010	0.050	0.099	0.148

I next simulate the data for $\Delta\mathcal{L}_{t,T}$ for the sample of length $T = 1000$ under $\mathbb{H}_0^{(3)}$ such that mean is time-varying:

$$\mathbb{H}_0^{(3)}: \qquad \Delta\mathcal{L}_{t,T} = \rho^0(t/T) + \rho^1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

and

$$\rho^0(u) = \sin(8\pi u), \qquad \rho^1(u) = 0 \quad \forall\ u \qquad \text{and} \qquad \sigma(u) = 1 \quad \forall\ u.$$

Under $\mathbb{H}_0^{(3)}$ the mean of $\Delta\mathcal{L}_t$ is time-varying. The mean performs four full sine cycles over the course of the sample, so that over the whole sample the overall mean is also zero by symmetry. For simplicity I set the variance to be constant throughout.

Table 3.5: Actual size versus nominal size of two-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(3)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.015	0.065	0.107	0.150
$B = 750$	0.014	0.061	0.105	0.145
$B = 1000$	0.012	0.060	0.105	0.145

Table 3.6: Actual size versus nominal size of one-sided $\mathcal{S}_T$ for $\mathbb{H}_0^{(3)}$.

Bootstrap size/ nominal α	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.10$	$\alpha = 0.15$
$B = 500$	0.010	0.043	0.087	0.141
$B = 750$	0.009	0.043	0.088	0.143
$B = 1000$	0.010	0.045	0.090	0.143

The above tables show that the actual size is very close to the nominal size for all levels and for all nulls under consideration. The results are stable regardless of the number of bootstrap replications B.

3.6.2 Test Statistics: Power

I start by replicating two alternatives from Giacomini and White (2006) that also constitute alternatives for my test. The first alternative simulates the loss differences $\Delta\mathcal{L}_t$ according to (3.19) such that $\rho = 0$ and $\mu = (0, 0.05, \cdots, 1)$. I fix the nominal size of the test to be 5%. Below I show the power curves when applying my one-sided and two-sided as well as Giacomini and White (2006) test.

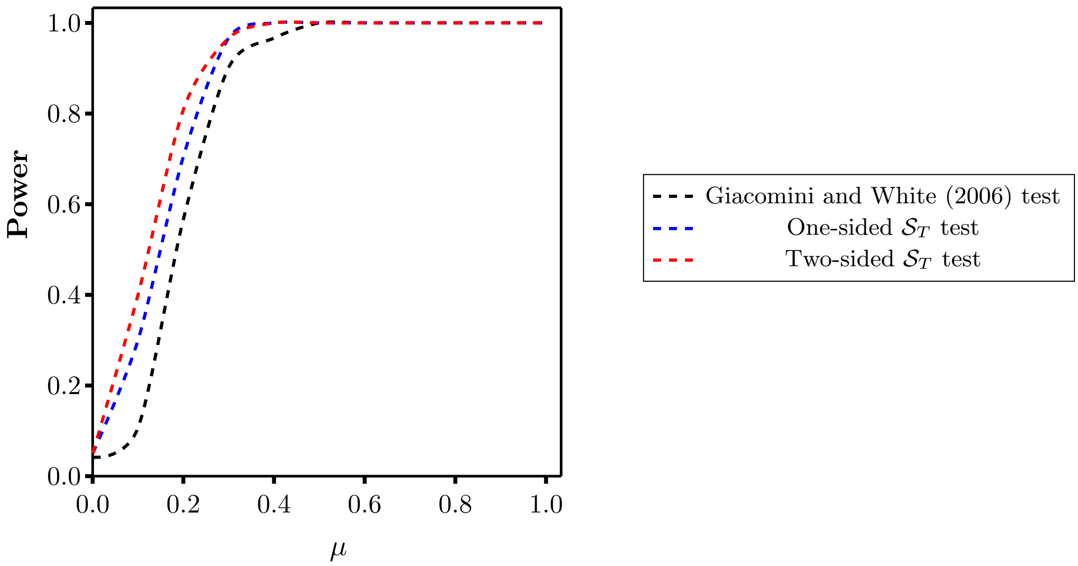


Figure 3.3: Power curves under alternative $\mathbb{H}_1^{(1)}$.

I next consider another alternative that I borrow from Giacomini and White (2006). In particular, I again generate the loss differences $\Delta\mathcal{L}_t$ according to (3.20), where I vary $d = \frac{\mu}{p(1-p)} = (0, 0.1, \cdots, 1)$. Note that d represents the difference in expected loss between two states. I apply my general test $\mathcal{S}_T'$ by setting the choice weighting functions to be the states of the world, i.e. I set $\phi_t = S_t$, conditional on the states of the world S_t. In this case (3.20) constitutes an alternative for my null as well. I plot the power curves below.

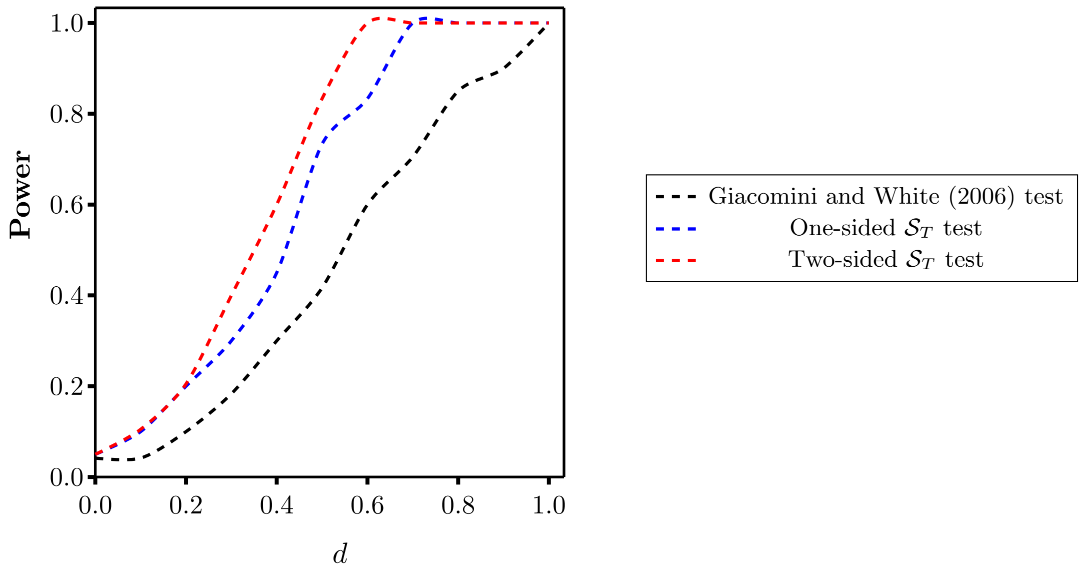


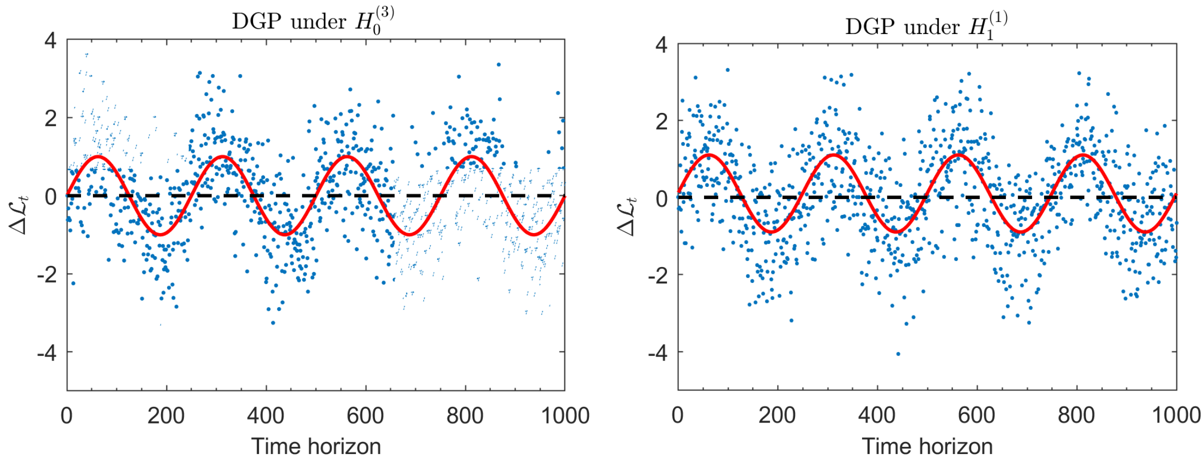
Figure 3.4: Power curves under alternative $\mathbb{H}_1^{(2)}$.

I now investigate the power of the test under several fixed alternatives that exhibit time variation of the mean/variance process. I deliberately design the set of these alternatives to be similar to the earlier time-varying null $\mathbb{H}_0^{(3)}$, however I add one additional feature that makes for a deviation from the null. Under the first alternative $\mathbb{H}_1^{(1)}$ I simulate the data as follows:

$$\Delta\mathcal{L}_{t,T} = \rho^0(t/T) + \rho^1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

where

$$\rho^0(u) = \sin(8\pi u) + 0.1, \qquad \rho^1(u) = 0 \quad \forall\, u \qquad \text{and} \qquad \sigma(u) = 1 \quad \forall\, u.$$



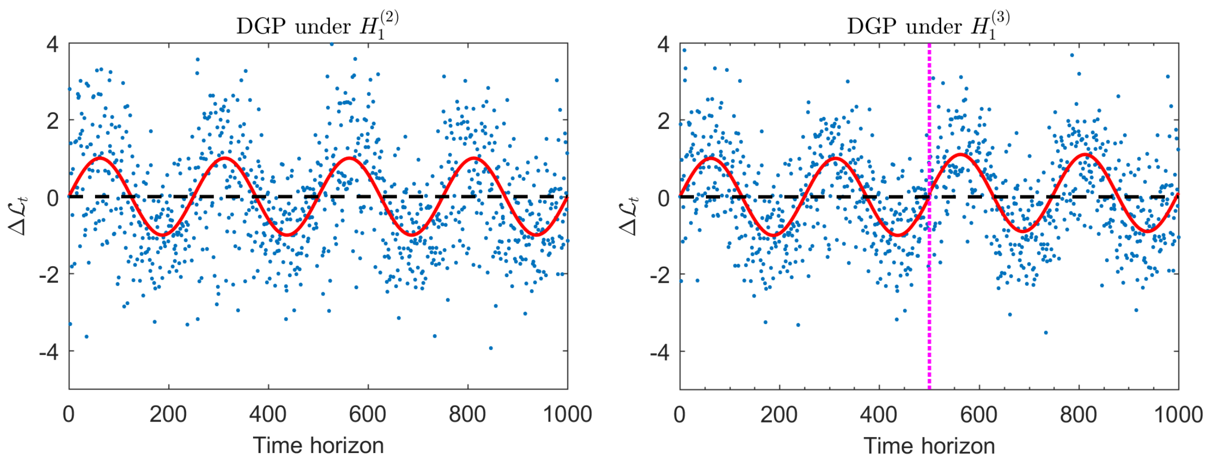


Figure 3.6: Data generating processes (DGP) under the null $\mathbb{H}_0^{(3)}$ as the corresponding alternatives $\mathbb{H}_1^{(2)}$, $\mathbb{H}_1^{(3)}$ and $\mathbb{H}_1^{(3)}$. The red lines represents the true mean function μ_t.

Under $\mathbb{H}_1^{(1)}$, I add a small intercept to the curve of the mean from the null. The deviation is hard to differentiate visually due to the variance around the mean, and the mean still goes above and below zero, with relative performance overtaking back and forth.

Under $\mathbb{H}_1^{(2)}$ I leave the mean the same as under the null and change the variance in a way that all upswings of the sine function are volatile and downswings are more volatile, more precisely:

$$\Delta\mathcal{L}_{t,T} = \rho^0(t/T) + \rho^1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

and where

$$\rho^0(u) = \sin(8\pi u), \qquad \rho^1(u) = 0 \quad \forall\ u,$$

and setting $w = T/8$, the local variance is given by

$$\sigma(u) = \begin{cases} 1 & \forall\ u \in [1+kw, (k+1)w] \quad \text{for} \quad k = 0, 2, 4, 6. \\ 1.5 & \forall\ u \in [1+kw, (k+1)w] \quad \text{for} \quad k = 1, 3, 5, 7. \end{cases}$$

Note that although the mean function under $\mathbb{H}_1^{(2)}$ is the same as under $\mathbb{H}_0$, due to the changes in the variance, the upper swings shall receive more weight as they are less volatile, while the opposite shall hold for the downswings. As the result, I expect the overall statistic to be positive, pointing towards the preference of model $\mathcal{B}$ versus the model $\mathcal{A}$.

Finally, I consider the alternative $\mathbb{H}_1^{(3)}$ that allows for a break in the mean function. In particular, under $\mathbb{H}_1^{(3)}$ I simulate the data as follows:

$$\Delta\mathcal{L}_{t,T} = \rho^0(t/T) + \rho^1(t/T)\Delta\mathcal{L}_{t-1,T} + \sigma(t/T)\varepsilon_t, \qquad \varepsilon_t \sim \mathcal{N}(0,1),$$

where $\rho^1(u) = 0 \quad \forall\, u \qquad \text{and} \qquad \sigma(u) = 1 \quad \forall\, u$, and

$$\rho^0(u) = \begin{cases} \sin(8\pi u) & \text{for } u \in [1, T/2], \\ \sin(8\pi u) + 0.1 & \text{for } u \in [T/2+1, T]. \end{cases}$$

This alternative highlights the ability for my test statistic to deal with breaks. Here the deviation to the null is smaller than the first alternative where the intercept added is throughout the whole sample.

Table 3.7: Mean of $\mathcal{S}_T$.

Alternative	$\mathbb{E}\left(\mathcal{S}_T\right)$
$\mathbb{H}_1^{(1)}$	3.08
$\mathbb{H}_1^{(2)}$	2.82
$\mathbb{H}_1^{(3)}$	1.40

Table 3.8: Power for different alternatives with two-sided null.

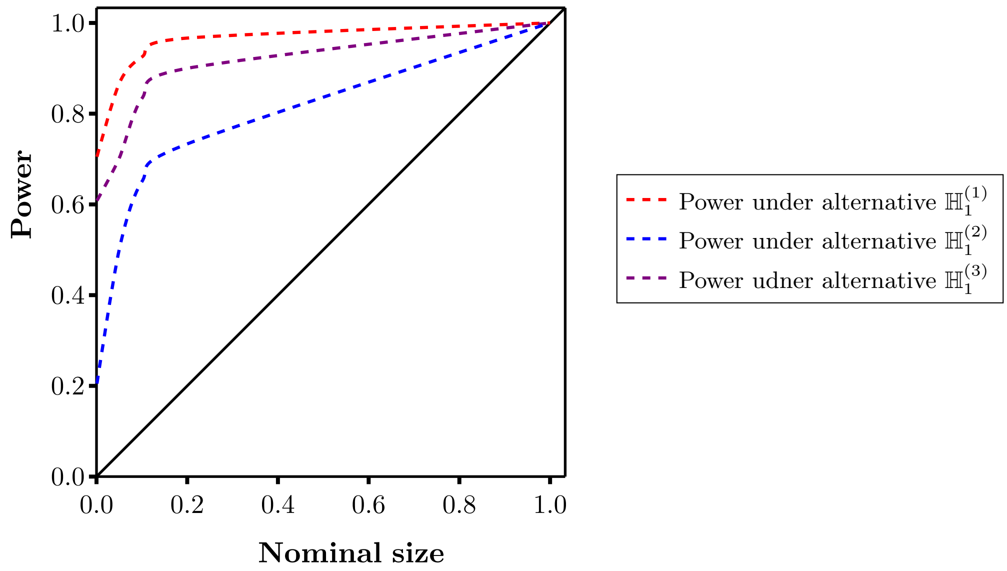
Nominal size	$\mathbb{H}_1^{(1)}$	$\mathbb{H}_1^{(2)}$	$\mathbb{H}_1^{(3)}$
$\alpha = 0.01$	0.75	0.60	0.24
$\alpha = 0.05$	0.88	0.64	0.38
$\alpha = 0.10$	0.94	0.75	0.50
$\alpha = 0.15$	0.97	0.85	0.57

Table 3.9: Mean of $\mathcal{S}_T$.

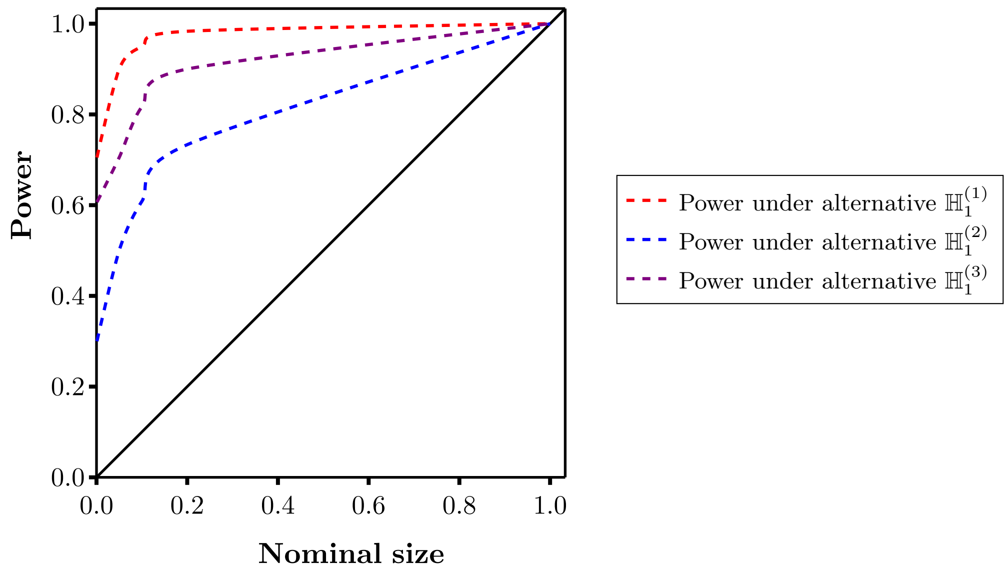
Alternative	$\mathbb{E}\left(\mathcal{S}_T\right)$
$\mathbb{H}_1^{(1)}$	3.08
$\mathbb{H}_1^{(2)}$	2.82
$\mathbb{H}_1^{(3)}$	1.40

Table 3.10: Power for different alternatives with one-sided null.

Nominal size	$\mathbb{H}_1^{(1)}$	$\mathbb{H}_1^{(2)}$	$\mathbb{H}_1^{(3)}$
$\alpha = 0.01$	0.76	0.66	0.28
$\alpha = 0.05$	0.90	0.74	0.48
$\alpha = 0.10$	0.96	0.83	0.62
$\alpha = 0.15$	0.98	0.90	0.72



(a) Two-sided test $\mathcal{S}_T$.



(b) One-sided test $\mathcal{S}_T$.

Figure 3.7: The figure plots the power curves for different alternatives. The dashed blue line depicts the power curve under $\mathbb{H}_1^{(3)}$, the dashed violet line depicts the power curve under $\mathbb{H}_1^{(2)}$, and the dashed red line depicts the power curve under $\mathbb{H}_1^{(1)}$.

Figure 3.7 shows that the test has a very good power at any nominal level and is capable of detecting relatively small deviations from the null. I finish this section with the following thought experiment. Assume that the true data generating process for $\Delta\mathcal{L}_t$ is indeed as under one of the considered alternatives $\mathbb{H}_1^{(1)}$, $\mathbb{H}_1^{(2)}$ or $\mathbb{H}_1^{(3)}$. Assume that the researcher applies any currently available test, e.g. Diebold and Mariano (1995) test or Giacomini and White (2006) test, to decide whether competing models have equal forecasting performance. As with any existing out-of-sample test the researcher would have to choose the splitting point. Table 3.11 displays the results of applying these tests as function of the cutoff point ρ, which is a fraction of the sample length used for forecast evaluation to the sample length used for the model estimation.

Table 3.11: Results of applying standard tests under different alternatives.

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_1^{(1)}$.

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.419	0.170	0.624	**0.002**	**0.042**	**0.033**	0.310	**0.040**	0.207	**0.026**
GW (2006)	**0.011**	**0.010**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_2^{(1)}$.

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.524	0.205	0.098	**0.031**	**0.012**	0.219	0.146	0.924	0.609	0.057
GW (2006)	0.586	0.206	0.100	**0.010**	**0.024**	**0.035**	**0.010**	**0**	**0**	**0**

Results when $\Delta\mathcal{L}_t$ is simulated according to $\mathbb{H}_3^{(1)}$.

p-value/Cutoff ρ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
DM (1995)	0.484	0.194	0.474	**0.014**	0.408	0.062	0.065	0.136	0.9496	**0.017**
GW (2006)	0.090	0.100	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**

Note: The cutoff point ρ is defined as a fraction of the evaluation to estimation samples, i.e. $\rho = T_2/T_1$, where T_2 is the length of the sample used for forecast evaluation and T_1 is the length of the sample used for estimation. The values in the table present the p-values from the test at the nominal level $\alpha = 5\%$. The p-values in bold indicate rejection at the nominal size $\alpha = 5\%$. DM abbreviates Diebold and Mariano (1995) test of equal predictive ability and GW abbreviates Giacomini and White (2006) test of conditional predictive ability with $h_t = [1, \Delta\mathcal{L}_{t-1}]'$.

Table 3.11 shows that the conclusion of the tests, especially the Diebold and Mariano (1995) test, can change depending on the splitting point when applied to my alternatives. The Giacomini and White (2006) test suffers less from the splitting point problem and with a reasonable estimation sample delivers consistent results. Interestingly, for many splitting points the Diebold and Mariano (1995) test does not reject the null of equal predictive ability, while the Giacomini and White (2006) test does reject the same null.[11] This is indicative of changing relative performance as we knew ex-ante, hence the existing methodology based on constant relative performance is inappropriate. I stress that the presented thought experiment is not a reflection on the tests as they were not designed to deal with the world of changing relative performance, but rather to highlight the dangers that the researcher runs into when applying existing tests

[11]Note that this result is not specific to the Diebold and Mariano (1995) test, but in fact to all existing tests that derive from it, including superior predictive ability tests, see White (2000), Hansen (2005) as well as Model Confidence Set test by Hansen et.al.(2011).

that rely on inappropriate assumption.

3.6.3 Sign Forecasting

In this section I assess how my methodology for sign forecasting, described in section 3.3, performs with a known data-generating process. In this case the true probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ is known. For simplicity, I choose the $\mathbb{H}_0^{(3)}$ as the true data generating process for $\Delta\mathcal{L}_t$ and forecast the probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, starting from $\underline{T} = 100$.

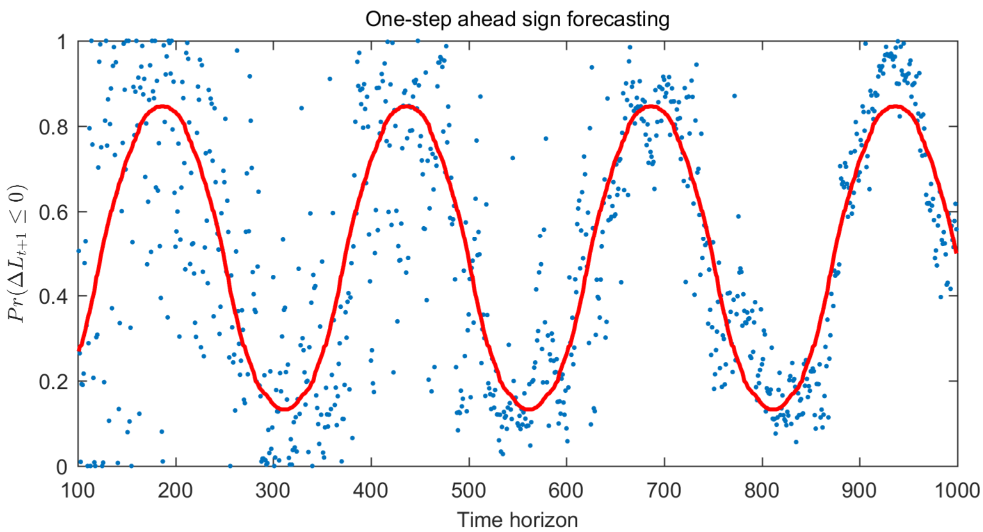


Figure 3.8: The red line plots the true probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ and the blue dots represent the estimate $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right).$

Figure 3.8 plots the true probability $Pr\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ against its estimate $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, where for each point on the curve the data up to $\underline{T}$ is used, where $\underline{T} = 100, \cdots, T$. Overall, the estimated probability is quite close to its true value and becomes more precise the more data is used for the original estimation. This happens primarily due to the c.d.f. of the error term $\widehat{\varepsilon}_t$ being better estimated towards the end of the sample as more data is used. At the final point in the sample, I forecast a probability of 0.3829 with a corresponding forecast interval of $[0.3520,\ 0.4200]$. Finally, applying the criterion, given in eq. (3.15), I get $\widehat{C} = -0.052$, which points to the fact that the estimated probability $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ is on average overestimated by approximately 5.2%.

3.7 Application

In this section I apply my proposed methodologies to the data. I first go back to the motivating example I presented in the introduction in Figure 3.1.

3.7.1 Motivating example in the Introduction

I consider the daily IBM returns spanning 03/01/2006-29/12/2016 and use two models to forecast daily variance: GARCH(1,1) model with Gaussian errors and GARCH(1,1) model with Student-t errors. The forecast loss is taken to be the squared error, see eq.(3.4) and constructed via the recursive scheme described in section 3.2. I compute the 5 minute realised volatility series from the data and it is taken to represent the "true" daily variance. Denote by $\widehat{\varepsilon}_t^{St}$ the forecast error when using the GARCH(1,1) model with Student-t errors for forecasting, and by $\widehat{\varepsilon}_t^{G}$ the forecast error from using the GARCH(1,1) model with Gaussian errors. I define $\Delta\mathcal{L}_t$ to be $\Delta\mathcal{L}_t := \left(\widehat{\varepsilon}_t^{St}\right)^2 - \left(\widehat{\varepsilon}_t^{G}\right)^2$. Once the $\{\Delta\mathcal{L}_t\}$ has been constructed, I apply my proposed two-step nonparametric procedure using AR(1) time-varying coefficient model (3.5) to estimate the corresponding time-varying mean and variance. Figure 3.9 depicts $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ and $\widehat{\tau}(t/T)$ calculated via eq.(3.9).

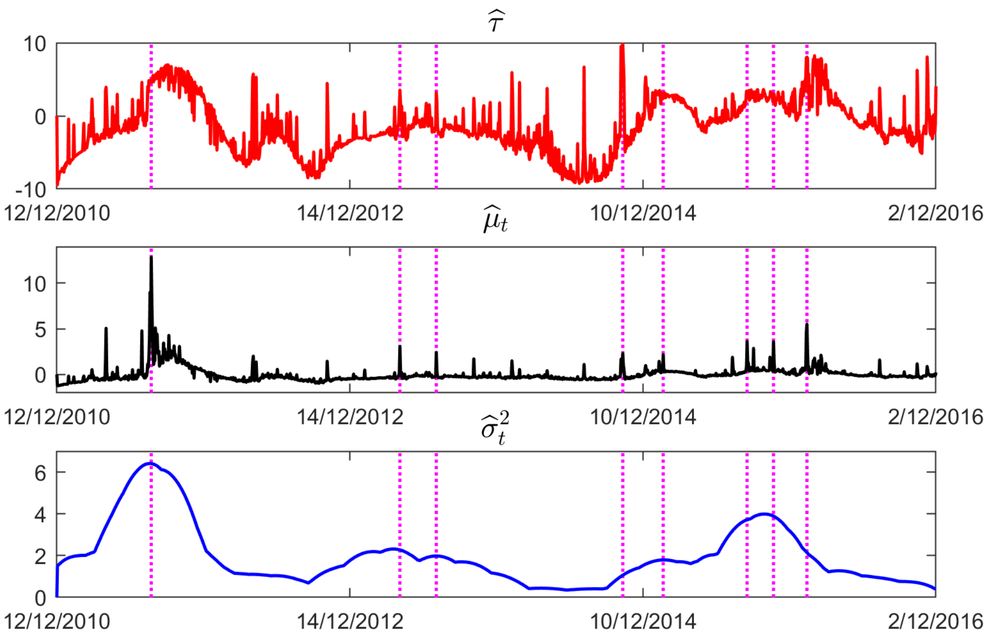


Figure 3.9: Plots of the estimates of $\widehat{\tau}$, $\widehat{\mu}_t$ and $\widehat{\sigma}_t^2$ for IBM data, 2006-2016, using squared error loss and recursive forecasting scheme.

Recall that each corresponding $\widehat{\mu}_t$ is weighted by the inverse of the standard error of $\widehat{\mu}_t(u)$. One can approximately take the weight to be $1/\widehat{\sigma}_t$. Hence whenever a spike occurs in the relative forecasting performance (represented by the violet dashed lines), the μ_t in those periods get down weighted. I next calculate the test statistic $\mathcal{S}_T$. Tables below show the critical values of two-sided and one-sided $\mathcal{S}_T$ test, specific to this application.

The value of the test statistics in this application example is $\mathcal{S}_T = -26.33$. Provided the critical values in Table 3.12, when the null of Equal Predictive Ability is tested, it is rejected at all levels of significance. Under the one-sided null of Superior Predictive Ability, the null is

Table 3.12: Critical values for $\mathcal{S}_T$.

Quantiles	0.005	0.025	0.05	0.075	0.85	0.90	0.925	0.95	0.975	0.99	0.995
Cr. values	-2.48	-1.90	-1.64	-1.37	1.00	1.43	1.61	1.72	2.10	2.60	2.77

Note: The critical values are calculated via the wild bootstrap and are specific to the application at hand.

not rejected for all significance levels, indicating that there is no evidence that the GARCH(1,1) model with normal errors is superior to the GARCH(1,1) model with Student-t. Under remark 4 the practitioner should default to choosing GARCH(1,1) with Student-t errors for forecasts. If supposing I test the opposite one-sided null, I find evidence that GARCH(1,1) with Student-t errors is superior to GARCH(1,1) with normal errors at all levels. Below I present the results of the pseudo out-of-sample sign forecasts.

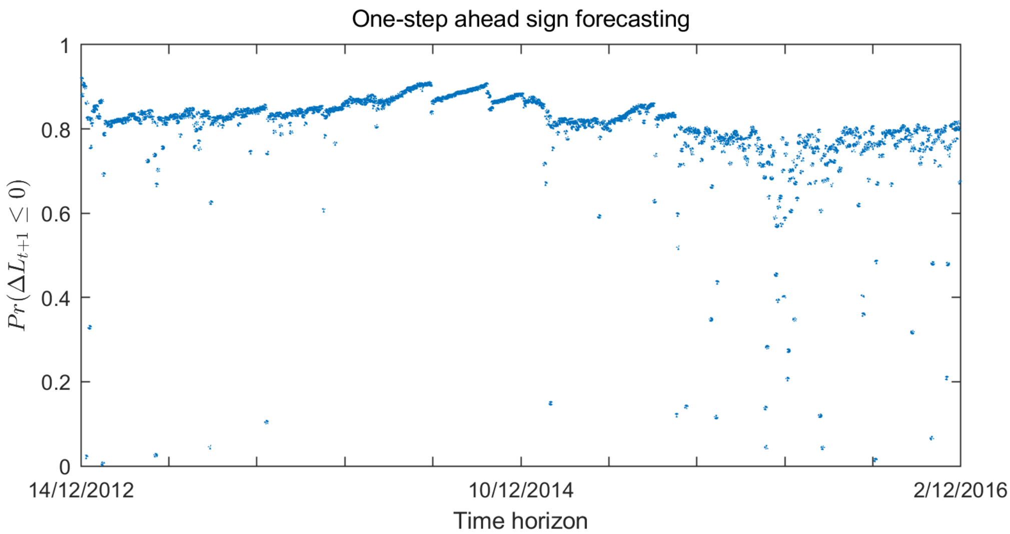


Figure 3.10: One-step ahead sign forecasting for the motivating example in the introduction.

We see that primarily, the probability of the GARCH(1,1) model with Student-t errors outperforming the GARCH(1,1) model with normal errors is relatively high for most points in time with a few exceptions. Finally applying the criterion given by eq. (3.15) I get the value $\widehat{C} = -0.033$, indicating that the forecasted probabilities are on average overestimated by 3.3%. At the final point in the sample, I forecast a probability of 0.3129 with a corresponding forecast interval of [0.2700, 0.3540]. Interestingly, this probability does not conclude that GARCH(1,1) with Student-t errors should be selected. This highlights the randomness inherent in forecasting next period probabilities. In this case, the two approaches would select different models for forecasting.

3.7.2 Comparing parameter-reduction methods

In this section I consider an application similar to that considered in Giacomini and White (2006). I consider the "balanced panel" of the dataset FRED-MD, consisting of 128 monthly economic time series measured over January, 1959 - August, 2017, and apply the same transformations to the original series, as documented in Appendix to the dataset.[12]. In particular, compared to Giacomini and White (2006), I extend their dataset to the end of August of 2017. I replace the sequential model examined in Giacomini and White (2006) with *lasso* to avoid multiple sequential testing. I then use several parameter-reduction methods, described below, to construct 1-month ahead forecasts of four US macroeconomic variables: two real variables - industrial production (abbreviated IP) and real personal income less transfers (abbreviated RPI); and two price indices: consumer price index (abbreviated CPI) and producer price index (abbreviated PPI).

All forecasting models project the $k-$step ahead variable of interest $\mathcal{Y}_{t+k}$ onto time t predictors $\mathcal{X}_t$ and lags of the variable of interest $\mathcal{Y}_t, \mathcal{Y}_{t-1}, \cdots$. I next describe the forecasting methods.

The full model for the k-step ahead forecast of the variable of interest $\mathcal{Y}_t$ is as follows:

$$\mathcal{Y}_{t+k} = \alpha + \beta\mathcal{X}_t + \gamma_1\mathcal{Y}_t + \gamma_2\mathcal{Y}_{t-1} + \cdots + \gamma_6\mathcal{Y}_{t-5} + \varepsilon_{t+k}, \tag{3.21}$$

where $\mathcal{X}_t$ contains all 135 predictors from the FRED-MD dataset. To overcome multicollinearity in $\mathcal{X}_t$, I follow Giacomini and White (2006) and replace the groups of variables in $\mathcal{X}_t$ whose correlation is greater that 0.98 with their average. The new X_t contains 120 predictors.

The first method considers the full model (3.21) and applies lasso to determine the relevant predictors. Denote by $Z_t = (X_t', \mathcal{Y}_t, \mathcal{Y}_{t-1}, \cdots, \mathcal{Y}_{t-5})'$, then lasso estimates the parameter vector $\theta = (\alpha, \beta', \gamma_1, \gamma_2, \cdots, \gamma_6)'$ by solving

$$\widehat{\theta} := \underset{\theta\in\mathbb{R}^d}{\arg\min}\left\{ \frac{1}{T}\sum_{t=1}^{T} \|\mathcal{Y}_t - \theta' Z_t\|_2^2 + \lambda\|\theta\|_1 \right\},$$

where d is the dimension of the parameter vector θ and $\|\cdot\|_2$ and $\|\cdot\|_1$ denotes the L_2- and L_1-norms respectively.

The next model I consider is the diffusion index method (abbreviated DI) that first uses principal component analysis to estimate j factors $\widehat{F}_t$ from the predictors $\mathcal{X}_t$ and then considers

[12]The FRED-MD dataset is collected and constantly updated by the Federal Reserve Bank of St. Louse and can be found online with the following link For the variables I consider in this chapter, the transformations are as follows: the first log difference for RPI and IP variables; and the second log difference for CPI and PPI variables.

the reduced model given by:

$$\mathcal{Y}_{t+k} = \alpha + \beta \widehat{F}_t + \gamma_1 \mathcal{Y}_t + \cdots + \gamma_p \mathcal{Y}_{t-p} + \varepsilon_{t+k},$$

where the lag length p is selected by BIC and the number of factors j is chosen by applying Onatski's (2009) test.

The Bayesian shrinkage method (abbreviated Bay) considers the full model (3.21) and applies Bayesian estimation with Normal-Gamma priors for the coefficients. Moreover, the Bayesian estimation is coupled with the use of the Elastic Net as a more stabilised version of lasso, see Zou and Hastie (2005), that also allows grouping effects. In particular, the Elastic Net estimator $\widehat{\theta}$ of the parameter vector $\theta = (\alpha, \beta', \gamma_1, \gamma_2, \cdots, \gamma_6)'$ is the solution of the following minimisation problem:

$$\widehat{\theta} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \frac{1}{2\sigma^2} \sum_{t=1}^{T} \left(\mathcal{Y}_t - \theta' Z_t\right)^2 + \lambda_1 \sum_{j=1}^{d} |\theta_j| + \lambda_2 \sum_{j=1}^{d} \theta_j^2,$$

where d is the dimension of the parameter vector θ and $Z_t = (X_t', \mathcal{Y}_t, \mathcal{Y}_{t-1}, \cdots, \mathcal{Y}_{t-5})'$. I follow Korobilis (2013) for setting the priors. In particular, the Bayesian prior for θ in the above penalised regression is

$$\pi(\theta|\sigma^2) \sim e^{-\frac{\lambda_1}{\sqrt{\sigma^2}} \sum\limits_{j=1}^{d} |\theta_j| - \frac{\lambda_2}{2\sigma^2} \sum\limits_{j=1}^{d} \theta_j^2},$$

and for the shrinkage parameter $\tau_j, j = 1, \cdots, d$ the hyperprior on τ_j^2 is given by

$$\pi(\tau_j^2|\lambda_1^2) \sim \text{Exponential}\left(\frac{\lambda_1^2}{2}\right), \qquad \text{for} \qquad j = 1, \cdots, d,$$

which leads to the prior of the parameter vector have the following diagonal covariance matrix:

$$V = \begin{pmatrix} \left(\tau_1^{-2} + \lambda_2\right)^{-1} & & & & \\ & \left(\tau_2^{-2} + \lambda_2\right)^{-1} & & & \\ & & \ddots & & \\ & & & \left(\tau_{d-1}^{-2} + \lambda_2\right)^{-1} & \\ & & & & \left(\tau_d^{-2} + \lambda_2\right)^{-1} \end{pmatrix}.$$

The benchmark methods are the autoregressive model (denoted by AR) given by:

$$\mathcal{Y}_{t+k} = \alpha + \gamma_1 \mathcal{Y}_t + \gamma_2 \mathcal{Y}_{t-1} + \cdots + \gamma_6 \mathcal{Y}_{t-5} + \varepsilon_{t+k},$$

where p is selected by BIC and $0 \leq p \leq 6$, and the random walk model (denoted by RW) in levels, corresponding to the forecasting model in differences $\mathcal{Y}_{t+k} = \alpha + \varepsilon_{t+k}$, which therefore captures just the unconditional mean of the variable of interest. I use the squared error as the loss function for evaluating the forecasts and construct the time series of losses for $k = 1$ month according the recursive scheme, described in Figure 3.2 with $\underline{T} = 100$.

The one-sided test $\mathcal{S}_T$ tests the null of SPA, against an alternative of inferior predictive ability. The loss differences are constructed from the loss of the model from the column model of the table, minus the loss of the model from the row model of the table. Therefore, a negative test statistic is indicative of SPA of the column model versus the row model, whereas a positive test statistic is indicative of the inferior predictive ability. The decision rule is to select the column model whenever there is not significant evidence to reject the null of SPA. Otherwise we select the row model. I highlight the cases when I reject the null of SPA in bold. I also identify that in general, the Bayesian model performs consistently poorly for all four variables. Conversely, the Random Walk model performs in general the best, except for forecasting personal income where it is insignificantly worse than the AR model.

My two-sided test allows to construct the overall ranking for models. The results are presented in Tables 3.13-3.14. Significant rejections in either direction of my null is highlighted again in bold. For each of the variables of interest, I obtain the following rankings:

- Personal income: AR $\geq$ RW$\geq$ DI $\geq$ Lasso $\geq$ Bay;
- Industrial production: RW $\geq$ Lasso $\geq$ AR $\geq$ DI $>$ Bay;
- Producer price index: RW$>$Lasso $\geq$ DI $\geq$ AR $\geq$ Bay;
- Consumer price index: RW $\geq$ Lasso $>$ DI $\geq$ AR $>$ Bay.

Where $\geq$ indicates an insignificant superior ranking and $>$ indicates a significant superior ranking. I remark that for all four variables, the Bayesian shrinkage method is consistently the worst in terms of its forecasting ability for $k = 1$ month, and significantly so for industrial production and the consumer price index. Conversely, the random walk model performs the best (followed by lasso), for all variables except personal income where the AR model is insignificantly ranked higher.

I now proceed to applying the sign forecasting methodology, described in section 3.3. For each of the variables I report the forecasted probability $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$ at the end of the sample as well as the associated forecast interval $[\widehat{FI}_l,\ \widehat{FI}_u]$. Also, for all variables I perform the pseudo out-of-sample sign forecasting exercise, starting with $\underline{T} = 100$, and report the value of my criterion $\widehat{C}$. Results are presented in Table 3.15. From the results in Table 3.15 I can infer the ranking of models based on forecasted next period next period performance. I do so in the following way. I say that model $\mathcal{A}$ outperforms model $\mathcal{B}$, denoted as $\mathcal{A} > \mathcal{B}$, if $\widehat{FI}_l > 0.5$

and $\mathcal{A} \geq \mathcal{B}$ if $0.5 \in [\widehat{FI}_l, \widehat{FI}_u]$ and $Pr\left(\Delta\mathcal{L}_{T+1}^{\mathcal{AB}} < 0\right) > 0.5$. Using the above described notation the ranking is as follows:

- Personal income: Lasso > AR > RW > DI > Bay;
- Industrial production: Lasso > AR > RW > DI > Bay;
- Producer price index: AR > Lasso > RW > DI > Bay;
- Consumer price index: AR > Lasso≥ RW > DI > Bay,

This ranking is to some degree similar to the ranking based on average past performance. In particular, at the bottom of the ranking we see that the Bayesian shrinkage method performs consistently the worst out of all models for all four variables. The diffusion index method (DI) performs also generally poorly, which is again consistent with the metric for past performance. One likely explanation of the poor performance of the DI method is the potential for overfitting, which translates into poor out-of-sample forecasting. The reason for the latter is that, in addition to the lags of the forecasted variable, the DI model also includes the k common factors extracted from the whole dataset, which might be irrelevant for forecasting in any particular period.

Interestingly, my sign forecasting approach to ranking indicates that the random walk model is not the best model for next period forecasting, as it is always dominated by either the lasso or the autoregressive model. In the case of the autoregressive model, it is likely because the autoregressive model can account for serial correlation in loss differences, which sometimes is the dominant feature in the data. Likewise, it appears that some of the time the lasso model is better able to capture next period performance than both the autoregressive model and the random walk model. We can also infer that my average performance metric, especially due to the weighting I employ, favors models that perform well consistently and with low variance over models that perform very well some of the time but not so well the other times. Hence the random walk model, as the conservative choice out of the selection of models, is often the best by my average performance metric. However forecasting one period ahead, we see it is the case that either or the autoregressive model will outperform. For the sake of brevity, I present the results for longer horizons, $k = 6$ and $k = 12$ months in Appendix C.

Table 3.13: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$.

		Personal Income					Industrial Production					Producer Price Index					Consumer Price Index				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.54					−2.13					−1.62					−2.92				
	p	0.936					0.980					0.948					0.998				
DI	$\mathcal{S}_T$	0.95	1.51				−0.64	2.11				−1.44	1.65				−4.21	2.94			
	p	0.172	0.086				0.768	**0.022**				0.938	**0.042**				1.00	**0.004**			
AR	$\mathcal{S}_T$	0.98	1.70	1.78			−0.46	2.17	0.08			−3.10	1.52	−1.79			−2.35	2.99	−1.46		
	p	0.164	**0.042**	**0.040**			0.666	**0.022**	0.468			0.998	0.082	0.958			0.992	**0.002**	0.936		
RW	$\mathcal{S}_T$	1.35	1.61	0.52	−1.76	–	0.91	2.06	0.41	1.48	–	4.75	1.50	2.58	2.96	–	1.70	3.02	3.65	2.90	–
	p	0.100	0.453	0.519	0.948	–	0.170	**0.028**	0.358	0.072	–	**0.000**	0.062	**0.008**	**0.004**	–	**0.046**	**0.000**	**0.000**	**0.005**	–

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (3.2), for horizon $k = 1$ month. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.54$ (indicating that Lasso is better) with the p-value of 0.936. The p-values in bold indicate rejection of the null (3.2) at the 5% level of significance.

Table 3.14: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.49					−2.04					−1.64					−3.07				
	p	0.152					**0.044**					0.172					**0.028**				
DI	$\mathcal{S}_T$	0.96	1.50				−0.64	2.20				−1.46	1.47				−4.32	2.99			
	p	0.304	0.124				0.524	**0.048**				0.148	0.124				**0.000**	**0.006**			
AR	$\mathcal{S}_T$	0.98	1.64	1.78			−0.46	2.13	0.08			−3.26	1.49	−1.47			−2.37	2.96	−1.40		
	p	0.296	0.132	0.052			0.628	**0.047**	0.92			**0.000**	0.108	0.068			**0.012**	**0.010**	0.152		
RW	$\mathcal{S}_T$	1.35	1.55	0.52	−1.80	–	0.86	2.02	0.58	1.75	–	4.96	1.54	2.59	3.08	–	1.63	2.98	3.52	3.02	–
	p	0.196	0.092	0.600	0.088	–	0.388	**0.048**	0.584	0.092	–	**0.000**	0.136	**0.008**	**0.000**	–	0.088	**0.007**	**0.000**	**0.000**	–

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds to the null of equal predictive ability, see eq. (3.1), for horizon $k = 1$ month. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.49$ (indicating that Lasso is better) with the p-value of 0.152. The p-values in bold indicate rejection of the null (3.1) at the 5% level of significance.

Table 3.15: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 1$ month.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.983					1.000					0.999					0.999				
	$\widehat{FI}_l$	0.976					1.000					0.997					0.998				
	$\widehat{FI}_u$	0.990					1.000					1.000					1.000				
	$\widehat{C}$	0.060					0.014					0.056					0.045				
DI	$\widehat{Pr}_{T+1}$	0.884	0.001				0.836	0.000				0.550	0.001				0.755	0.002			
	$\widehat{FI}_l$	0.839	0.000				0.809	0.000				0.520	0.000				0.716	0.001			
	$\widehat{FI}_u$	0.905	0.005				0.849	0.000				0.578	0.006				0.771	0.003			
	$\widehat{C}$	-0.024	-0.078				-0.032	-0.014				0.067	-0.068				0.071	-0.032			
AR	$\widehat{Pr}_{T+1}$	0.954	0.001	0.089			0.537	0.000	0.212			0.319	0.001	0.570			0.436	0.001	0.412		
	$\widehat{FI}_l$	0.912	0.000	0.064			0.510	0.000	0.195			0.281	0.000	0.550			0.386	0.000	0.385		
	$\widehat{FI}_u$	0.986	0.002	0.126			0.560	0.000	0.246			0.344	0.004	0.600			0.460	0.002	0.456		
	$\widehat{C}$	-0.018	-0.105	-0.049			-0.090	-0.014	-0.017			0.037	-0.049	-0.034			0.058	-0.039	-0.025		
RW	$\widehat{Pr}_{T+1}$	0.829	0.002	0.078	0.816		0.660	0.000	0.181	0.624		0.944	0.002	0.380	0.669		0.978	0.001	0.162	0.523	
	$\widehat{FI}_l$	0.802	0.001	0.066	0.790	–	0.628	0.000	0.160	0.597	–	0.917	0.001	0.348	0.620	–	0.920	0.000	0.141	0.482	–
	$\widehat{FI}_u$	0.861	0.004	0.115	0.839	–	0.690	0.000	0.214	0.654	–	0.954	0.004	0.413	0.706	–	0.997	0.001	0.201	0.572	–
	$\widehat{C}$	-0.026	-0.085	-0.023	0.041		-0.015	-0.015	-0.011	0.001		-0.048	-0.069	-0.053	-0.034		-0.037	-0.043	-0.061	-0.036	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 1$ month. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(3.15). The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.983$ with the corresponding forecast interval $[0.976,\ 0.990]$.

3.8 Concluding remarks

In this chapter I address the issue of forecast evaluation and forecast selection in unstable environments. Existing out-of-sample tests often suffer from low power, and in unstable environments they can generate spurious and potentially misleading results. I address the possibility of unstable environments explicitly, and provide two methods by which to inform the selection of models for future forecasts. Importantly, my new methodology is no longer reliant on a sample splitting point, which is directly connected to the two limitations of the existing out-of-sample tests. I demonstrate that my methodology performs well across a variety of applications, and my test has high power against a range of fixed and local alternatives.

References

Bollerslev, T., Quaedvlieg, R., and Patton A.J., 2016, Exploiting the errors: a simple approach for improved volatility forecasting, *Journal of Econometrics*, 192, 1-18.

Bradley, R.C., 2005, Basic properties of strong mixing conditions. A survey and some open questions, *Probability Surveys*, 2, 107-144.

Cai, Z., 2007, Trending time-varying coefficient time series models with serially correlated errors,*Journal of Econometrics*, 136, 163-288.

Cai, Z., Fan, J., and Li, R.Z., 2000, Efficient estimation and inferences for varying-coefficient models, *Journal of American Statistical Association*, 95, 888-902.

Cai, Z., Fan, J., and Yao, Q., 2000, Functional-coefficient regression models for nonlinear time series, *Journal of the American Statistical Association*, 95 (451), 941-956.

Clark, T.E, and McCracken, M.W., 2001, Tests of equal forecast accuracy and encompassing for nested models, *Journal of Econometrics*, 105(1), 85-110.

Clark, T.E, and McCracken, M.W., 2005, Evaluating direct multistep forecasts, *Econometrics Reviews*, 24, 369-404.

Diebold, F.X., 2013, Comparing predictive accuracy, twenty years later: a personal perspective on the use and abuse of diebold-mariano tests, *Journal of Business and Economic Statistics*.

Diebold, F.X., and Mariano R.S., 1995, Comparing predictive accuracy, *Journal of Business and Economic Statistics*, 13, 253-263.

Fan J., and Yao Q., 1998, Efficient estimation of conditional variance functions in stochastic regression, *Biometrika*, 85(3), 645-660.

Giacomini R., and Rossi B., 2009, Detecting and predicting forecast breakdowns, *The Review of Economic Studies* 76(2), 669-705.

Giacomini, R. and Rossi, B., 2010, Forecast comparisons in unstable environments, *Journal of Applied Econometrics*, 25(4), 595-620.

Giacomini R., and White H., 2006, Test of conditional predictive ability, *Econometrica* 74(6), 1545-1578.

Corradi V., and Swanson, N.R., 2007, Nonparametric bootstrap procedures for predictive inference based on the recursive estimation schemes, *International Economic Review*, 48, 67-109.

Hall, P., and Heyde, C.C, 1980, *Martingale limit theory and its applications*, Academic Press.

Hansen B.E., 2008, Least squares forecast averaging," *Journal of Econometrics*, 146, 342-350.

Hansen B.E., 2007, Least squares model averaging," *Econometrica*, 75, 1175-1189.

Hansen, P.R., 2005, A test for superior predictive ability, *Journal of Business & Economic Statistics*, 23(4), 365-380.

Hansen, P.R., 2010, A winner's curse for econometric models: on the joint distribution of in-sample fit and out-of-sample fit and its implications for model selection, *Stanford University Working paper.*

Hansen P.R., and Timmermann, A., 2015, Discussion of comparing predictive accuracy, twenty years later by Francis X. Diebold, *Journal of Business and Economics Statistics*, 33, 17-21.

Hansen P. R., and Timmermann, A., 2010, Choice of sample split in out-of-sample forecast evaluation, Working paper.

Hansen P.R., Lunde, A., and Nason, J.M., 2011, Model confidence set, *Econometrica*, 79(2), 453-497.

Hirano K., and Wright J.H., 2017, Forecasting with model uncertainty: representations and risk reduction, *Econometrica*, 85 (2), 617-643.

Inoue, A., and Kilian L., 2004, In-sample or out-of-sample tests of predictability: which one should we use?, *Econometric Reviews*, 23(4), 371-402.

Inoue, A. and Rossi, B., 2012, Out-of-sample tests robust to the choice of the window size, *Journal of Business and Economics Statistics* 30(3), 432-453.

Korobilis, D., 2013, Hierarchical shrinkage priors for dynamic regressions with many predictors, *International Journal of Forecasting*, 29, 43-59.

Kristensen, D., 2009, Uniform convergence rates of kernel estimators with heterogeneous dependent data, *Econometric Theory*, 25(5), 1433-1445.

Kristensen, D., 2012, Non-parametric detection and estimation of structural change, *Econometrics Journal*, 15, 420-461.

Li, J., and Patton, A.J., 2017, Asymptotic inference about predictive accuracy using high-frequency, *Duke University Working paper.*

Li, Q., and Racine J.S., 2007, *Nonparametric Econometrics: Theory and Practice*, Princeton University Press.

McCracken, M.W., 2000, Robust out-of-sample inference, *Journal of Econometrics*, 99, 195-223.

McCracken, M.W., 2007, Asymptotics for out-of-sample tests of granger causality, *Journal of Econometrics*, 140, 719-752.

Mincer, J., and Zarnowitz, V., 1969, The evaluation of economic forecasts, in *Economic Forecasts and Expectations*, ed. J. Mincer, New York: National Bureau of Economic Research.

Onatski A., 2009, Testing hypotheses about the number of factors in large factor models, *Econometrica*, 77 (5), 1447-1479.

Orbe, S., Ferreira, E., and Rodriguez-Poo, J., 2005, Nonparametric estimation of time varying parameters under the shape restrictions, *Journal of Econometrics*, 126, 53-77.

Pollard, D., 1984, *Convergence of stochastic processes*, Springer, New York.

Stinchcombe, M.B., and White, H., 1998, Consistent specification testing with nuisance parameters present only under the alternative, *Econometric Theory*, 14, 295-325.

Vogt, M., 2012, Nonparametric regression for locally stationary time series, *Annals of Statistics*, 40, 2601-2633.

West, K.D., 1996, Asymptotic inference about predictive ability, *Econometrica*, 64, 1067-1084.

White, H., 2000, A reality check for data snooping, *Econometrica*, 68(5), 1097-1126.

Zou, H., and Hastie, T., 2005, Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society, Series B*, 67, 301-320.

3.9 Appendix A.

Assume the true data generating process for $\{y_t\}_{t=1}^T$ follows an AR(1) process:

$$y_t = \rho y_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim i.i.d.(0, \sigma^2), \qquad |\rho| < 1,$$

where ε_t is a m.d.s. Assume one uses two simple models to forecast y_t one-step ahead:

- Model $\mathcal{A}$ uses $\widehat{y}_{t+1|t} = 0$ for all $t = 1, \cdots, T$ as a forecast for y_{t+1};
- Model $\mathcal{B}$ uses $\widehat{y}_{t+1|t} = 0.1$ for all $t = 1, \cdots, T$ as a forecast for y_{t+1};

Assume also that the forecaster uses the mean squared error (MSE) to assess the quality of the forecasts, i.e.

$$\mathcal{L}_t^{\mathcal{A}} = \mathbb{E}\left[\left(y_{t+1} - \widehat{y}_{t+1|t}\right)^2 |\mathcal{F}_t\right] = \rho^2 y_t^2 + \sigma^2,$$

and

$$\mathcal{L}_t^{\mathcal{B}} = \mathbb{E}\left[\left(y_{t+1} - \widehat{y}_{t+1|t}\right)^2 |\mathcal{F}_t\right] = \rho^2 y_t^2 + \sigma^2 - 0.2\rho y_t + 0.01,$$

and therefore

$$\Delta\mathcal{L}_t^{\mathcal{AB}} = \mathcal{L}_t^{\mathcal{A}} - \mathcal{L}_t^{\mathcal{B}} = 0.01 - 0.2\rho y_t. \tag{3.22}$$

From eq.(3.22) it then follows that

$$\begin{cases} \Delta\mathcal{L}_t^{\mathcal{AB}} \leq 0 & \text{if} \quad y_t > 0.05/\rho, \\ \Delta\mathcal{L}_t^{\mathcal{AB}} > 0 & \text{if} \quad y_t < 0.05/\rho. \end{cases}$$

3.10 Appendix B.

This Appendix presents proofs of the theoretical results. Throughout the proofs for brevity of notation I will drop the second subscript, i.e. for any variable $X_{t,T}$ I will just write X_t.

Proof of Theorem 1.

The proof is based on that of Kristensen (2012) by extending the notation to accommodate the locally linear estimator of $\theta(t/T)$. I first lay out the notation used in establishing Theorem 1, since the rest of subsequent theory will use the same notation. Recall that I model $\Delta\mathcal{L}_t$ as the following varying-coefficient model:

$$\Delta\mathcal{L}_t = \rho_0(t/T) + \sum_{t=1}^{d} \rho_j(t/T)\Delta\mathcal{L}_{t-j} + \xi_t, \qquad \xi_t = \sigma(t/T)\varepsilon_t.$$

I further make use of the following notation: $\mathbb{X}_t = (1, \Delta\mathcal{L}_{t-1}, \cdots, \Delta\mathcal{L}_{t-d})^T$ and $\rho(t/T) = (\rho_0(t/T), \rho_1(t/T), \cdots, \rho_d(t/T))^T$. In addition, using the first-order Taylor approximation I have:

$$\rho_j(t/T) = a_j + b_j\,(t/T - u) + o\,(t/T - u)\,, \qquad 0 \leq j \leq d,$$

where $a_j = \rho_j(u)$ and $b_j = \dot{\rho}_j(u)$. Denote further by $\mathbb{Z}_t = \left(\mathbb{X}_t^T, \mathbb{X}_t^T\,(t/T - u)\right)^T$ and $\theta = \theta(u) = \left(\rho^T(u), \dot{\rho}^T(u)\right)^T$. Then the locally weighted least squares is

$$\widehat{\theta}(u) = \arg\min_{\theta} \sum_{t=1}^{T} K_{h_1}\,(t/T - u)\left(\Delta\mathcal{L}_t - \mathbb{Z}_t^T\theta\right)^2. \tag{3.23}$$

Minimising (3.23) with respect to θ provides the local linear estimator of $\rho_j(u)$, denoted by $\widehat{\rho}_j(u)$, which are the first $(d+1)$ elements of $\widehat{\theta}$ and the local linear estimator of the derivatives of $\rho_j(u)$, denoted by $\widehat{\dot{\rho}}_j(u)$, which are the last $(d+1)$ elements of $\widehat{\theta}$. It is straightforward to show that

$$\widehat{\theta}(u) = \begin{bmatrix} \Sigma_{T,0}(u) & \Sigma_{T,1}(u) \\ \Sigma_{T,1}(u) & \Sigma_{T,2}(u) \end{bmatrix}^{-1} \begin{pmatrix} W_{T,0}(u) \\ W_{T,1}(u) \end{pmatrix} = \Sigma_T^{-1}(u)W_T(u),$$

where

$$\Sigma_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\,(t/T - u)\left(t/T - u\right)^m \mathbb{X}_t\mathbb{X}_t^T, \qquad \text{for} \qquad m = 0, 1, 2,$$

and

$$W_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\,(t/T - u)\left(t/T - u\right)^m \mathbb{X}_t\Delta\mathcal{L}_t, \qquad \text{for} \qquad m = 0, 1.$$

Recall that $\Omega_{t,T} \equiv \mathbb{E}\left[\mathbb{X}_t\mathbb{X}_t^T\right] = \Omega(t/T) + o(1)$ and $H = \operatorname{diag}\left(I_{d+1}, h_1 I_{d+1}\right)$ with I_{d+1} being the $(d+1)\times(d+1)$ identity matrix. In addition, I define the following quantities:

$$V(t/T,u) = \rho(t/T) - \left\{\rho(u) + \dot{\rho}(u)(t/T-u) + \frac{1}{2}\ddot{\rho}(u)(t/T-u)^2\right\} \tag{3.24}$$

$$\widetilde{W}_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m \mathbb{X}_t\xi_t, \qquad \mathbb{B}_{T,m}(u) = \frac{1}{2}\Sigma_{T,m+2}(u)\ddot{\rho}(u),$$

and

$$\mathbb{R}_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^m \mathbb{X}_t\mathbb{X}_t^T V(t/T,u). \tag{3.25}$$

With the above notation I can rewrite $W_{T,m}$ for $m=0,1$ as follows:

$$W_{T,m} = \Sigma_{T,m}(u)\rho(u) + \Sigma_{T,m+1}(u)\dot{\rho}(u) + \widetilde{W}_{T,m}(u) + \mathbb{B}_{T,m}(u) + \mathbb{R}_{T,m}(u).$$

Then it holds that

$$\widehat{\theta}(u) - \theta(u) - \Sigma_T^{-1}(u)\mathbb{B}_T(u) - \Sigma_T^{-1}(u)\mathbb{R}_T(u) = \Sigma_T^{-1}(u)\widetilde{W}_T(u). \tag{3.26}$$

To establish the proof of Theorem 1 it remains to show the following results:

(C1) $\sup_{\theta\in I_{h_1}}\|\widehat{\theta}(u)-\theta(u)\| = O_p\left(\sqrt{\frac{\log T}{Th_1}} + h_1^2\right)$, where $I_{h_1} := [Ch_1, 1-Ch_1]$ and $C1>0$ such that $Ch_1\to 0$ and $1/C_1 \to 0$,

(C2) $H\{\widehat{\theta}(u)-\theta(u)\} = h_1^2\mathbb{B}_1(u) + o_p(1)$, and

(C3) $\sqrt{Th_1}H^{-1}\widetilde{W}_T(u) \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}_\theta(\theta)\right)$, where $\mathbb{B}_1(u)$ and $\mathbb{V}_\theta(u)$ are given in Theorem 1.

Proof of (C1) and (C2). In light of Assumptions $A1$ and $A2$ the process $\Delta\mathcal{L}_t$ is strongly mixing and $\sup_{t,T}\mathbb{E}\left[|\Delta\mathcal{L}_t|^s\right] < \infty$ for $s>4$, see Orbe et. al.(2005) Lemma A4 for a proof of this result. Once the mixing condition of $\Delta\mathcal{L}_t$ is established as well as the finiteness of its moments, in conjunction with Assumptions $A1-A3$ we now satisfy e.g. Assumptions $A1-A6$ for Theorem 1 in Kristensen (2009) or Assumptions $K1-K3$ for Theorem 4.1 in Vogt (2012) to conclude that

$$\sup_{\theta\in I_{h_1}}\|\widehat{\theta}(u)-\theta(u)\| = O_p\left(\sqrt{\frac{\log T}{Th_1}} + h_1^2\right),$$

where $I_{h_1} := [Ch_1, 1-Ch_1]$ and $C1>0$ such that $Ch_1 \to 0$ and $1/C_1\to 0$. To establish (C2) note that $V(t/T,u) = o_p(h_1^2)$ and we therefore can ignore this term and focus on $\mathbb{B}_{T.m}(u)$ for $m=0,1$.

First consider $\Sigma_{T,m}(u)$ terms. Using Riemann sum approximation of an integral the following holds:

$$h_1^{-m}\mathbb{E}\left[\Sigma_{T,m}(u)\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{X}_t\mathbb{X}_t^T K_{h_1}(t/T-u)\left(\frac{t/T-u}{h_1}\right)^m\right] =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_1}(t/T-u)\left(\frac{t/T-u}{h_1}\right)^m \Omega(t/T) + o(1) = \int_{-1}^{1} y^m K(y)\Omega(u+yh_1)dy + o(1) =$$

$$= \lambda_m\Omega(u) + o(1).$$

Therefore it holds that

$$h_1^{-m}\Sigma_{T,m}(u) = \lambda_m\Omega(u)\{1+o_p\}, \qquad \text{i.e.} \qquad h_1^{-m}\Sigma_{T,m}(u) \xrightarrow{p} \lambda_m\Omega(u), \tag{3.27}$$

and

$$H^{-1}\mathbb{B}_T(u) = \frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\ 0\end{pmatrix}\otimes\ddot{\rho}(u) + o_p(h_1^2), \tag{3.28}$$

We can therefore re-write (3.26) as follows:

$$\sqrt{Th_1}\left(H\{\widehat{\theta}(u)-\theta(u)\} - \frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\ 0\end{pmatrix}\otimes\ddot{\rho}(u) + o_p(h_1^2)\right) =$$

$$= \left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}\sqrt{Th_1}H^{-1}\widetilde{W}_T(u) = \Sigma(u)^{-1}\sqrt{Th_1}H^{-1}\widetilde{W}_T(u)\{1+o_p(1)\} = O_p(1), \tag{3.29}$$

where in light of (3.27)

$$\Sigma(u) = \begin{pmatrix}\Omega(u) & 0\\ 0 & \lambda_2\Omega(u)\end{pmatrix}.$$

Proof of (C3). Define $Y_{t,T}^m := h_1^{-m-1/2}K\left(\frac{t/T-u}{h_1}\right)(t/T-u)^m\mathbb{X}_t\xi_t$, which is a martingale difference sequence w.r.t. $\mathcal{F}_t = \sigma(\Delta\mathcal{L}_t, \varepsilon_t, \Delta\mathcal{L}_{t-1}, \varepsilon_{t-1}, \cdots)$. To complete the proof of (C3) it suffices to verify Lemma B.13 in Kristensen (2012) for $Y_{t,T}^m$, $m = 0, 1$. In particular, as $Th_1 \to \infty$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[Y_{t,T}^m(Y_{t,T}^m)'\right] = \frac{1}{Th_1}\sum_{t=1}^{T}K\left(\frac{t/T-u}{h_1}\right)^2\left(\frac{t/T-u}{h_1}\right)^{2m}\Omega(t/T)\sigma(t/T)^2 + o(1) =$$

$$= \frac{1}{h_1}\int_{-1}^{1}K^2\left(\frac{y-u}{h_1}\right)\left(\frac{y-u}{h_1}\right)^{2m}\sigma^2(y)\Omega(y)dy + o(1) = \nu_{2m}\sigma^2(u)\Omega(u) + o(1),$$

and

$$\frac{1}{T^{1+\delta/2}}\sum_{t=1}^{T}\mathbb{E}\left[\|Y_{t,T}^{m}\|^{2+\delta}\right]=\frac{1}{(Th_1)^{1+\delta/2}}\sum_{t=1}^{T}K^{2+\delta}\left(\frac{t/T-u}{h_1}\right)\left(\frac{t/T-u}{h_1}\right)^{(2+\delta)m}\mathbb{E}\left[\|\mathbb{X}_t\|^{2+\delta}|\varepsilon_t|^{2+\delta}\right]=$$

$$=\frac{C}{(Th_1)^{\delta/2}}\sigma^{2+\delta}(u)\int_{-1}^{1}K^{2+\delta}(y)y^{m(2+\delta)}dy=o(1).$$

Therefore $\sqrt{Th_1}H^{-1}\widetilde{W}_T\xrightarrow{d}\mathcal{N}\left(0,\ \Xi[u]\right)$, where

$$\Xi(u)=\begin{pmatrix}\nu_0\sigma^2(u)\Omega(u) & 0\\ 0 & \nu_2\sigma^2(u)\Omega(u)\end{pmatrix},$$

Combining all of the above with (3.29), it then follows that

$$\sqrt{Th_1}\left(H\{\widehat{\theta}(u)-\theta(u)\}-\frac{h_1^2}{2}\begin{pmatrix}\lambda_2\Omega(u)\\ 0\end{pmatrix}\otimes\ddot{\rho}(u)+o_p(h_1^2)\right)\xrightarrow{d}\mathcal{N}\left(0,\ \Sigma(u)^{-1}\Xi(u)\Sigma(u)^{-1}\right),$$

which completes the proof of Theorem 1. $\blacksquare$.

Proof of Theorem 2.

Define the estimated errors $\widehat{\xi}_t$:

$$\widehat{\xi}_t=\Delta\mathcal{L}_t-\mathbb{Z}_t^T\widehat{\theta}(t/T)=\mathbb{Z}_t^T\theta(t/T)+\xi_t-\mathbb{Z}_t^T\widehat{\theta}(t/T)=\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}+\xi_t.$$

Running the local constant nonparametric regression of $\widehat{\xi}_t^2$ on rescaled time I get:

$$\widehat{\sigma}^2(u)=\frac{\frac{1}{T}\sum\limits_{t=1}^{T}K_{h_2}\left(t/T-u\right)\widehat{\xi}_t^2}{\frac{1}{T}\sum\limits_{t=1}^{T}K_{h_2}\left(t/T-u\right)}.$$

Note also that

$$\widehat{\xi}_t^2=\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t+\xi_t^2+2\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}\xi_t,$$

and therefore

$$\widehat{\xi}_t^2-\xi_t^2=\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}+2\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t\xi_t=$$

$$=\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}^T\mathbb{Z}_t\mathbb{Z}_t^T\left\{\theta(t/T)-\widehat{\theta}(t/T)\right\}-2\left\{\widehat{\theta}(t/T)-\theta(t/T)\right\}^T\mathbb{Z}_t\xi_t.$$

I therefore can write:

$$
\begin{aligned}
\widehat{\sigma}^2(u) - \sigma^2(u) &= \left\{ \frac{1}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - h\right) \left[\widehat{\xi}_t^2 - \sigma^2(u) \right] \right\} \{1 + o_p(1)\} = \\
&= \frac{1}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \left[\left(\mathbb{Z}_t^T \left\{ \theta(t/T) - \widehat{\theta}(t/T) \right\} \left\{ \theta(t/T) - \widehat{\theta}(t/T) \right\}^T \mathbb{Z}_t + \sigma^2(t/T)\varepsilon_t^2 + \right.\right. \\
&\qquad \left.\left. - 2\mathbb{Z}_t^T \left\{ \widehat{\theta}(t/T) - \theta(t/T) \right\} \sigma(t/T)\varepsilon_t \right) - \sigma^2(u) \right] = \\
&= \frac{1}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \left[\sigma^2(t/T) - \sigma^2(u) \right] + \frac{1}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \sigma^2(t/T) \left\{ \varepsilon_t^2 - 1 \right\} + \\
&\qquad - \frac{2}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \mathbb{Z}_t^T \{ \widehat{\theta}(t/T) - \theta(t/T) \} \sigma(t/T)\varepsilon_t + \\
&\qquad + \frac{1}{T\widehat{f}(u)} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \mathbb{Z}_t^T \left\{ \theta(t/T) - \widehat{\theta}(t/T) \right\} \left\{ \theta(t/T) - \widehat{\theta}(t/T) \right\}^T \mathbb{Z}_t.
\end{aligned}
$$

To make it easier to work with the above expression, I write it as follows:

$$
\widehat{\sigma}^2(u) - \sigma^2(u) = \frac{1}{\widehat{f}(u)} \Big(\widehat{\mathcal{I}}_1(u) + \widehat{\mathcal{I}}_2(u) + \widehat{\mathcal{I}}_3(u) + \widehat{\mathcal{I}}_4(u) - \sigma^2(u)\widehat{f}(u) \Big),
$$

with

$$
\widehat{f}(u) = \frac{1}{T} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right),
$$

$$
\widehat{\mathcal{I}}_1(u) = \frac{1}{T} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \sigma^2(t/T),
$$

$$
\widehat{\mathcal{I}}_2(u) = \frac{1}{T} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \sigma^2(t/T) \{\varepsilon_t^2 - 1\},
$$

$$
\widehat{\mathcal{I}}_3(u) = -\frac{2}{T} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \mathbb{Z}_t^T \{ \widehat{\theta}(t/T) - \theta(t/T) \} \sigma(t/T)\varepsilon_t
$$

and

$$
\widehat{\mathcal{I}}_4(u) = \frac{1}{T} \sum_{t=1}^{T} K_{h_2}\left(t/T - u\right) \mathbb{Z}_t^T \{ \widehat{\theta}(t/T) - \theta(t/T) \} \{ \widehat{\theta}(t/T) - \theta(t/T) \}^T \mathbb{Z}_t.
$$

I start by deriving some preliminary results for the expressions above:

i) By applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and $W_{t,T} = 1$ I arrive at the following result:

$$
\sup_{u \in I_{h_2}} \left| \widehat{f}(u) - f(u) \right| = o_p(1). \tag{3.30}
$$

Moreover, (3.30) together with an extra condition that $\inf_{u\in[0,1]} f(u) > 0$ implies that

$$\sup_{u\in(0,1)} \widehat{f}(u)^{-1} = O_p(1).$$

ii) By applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and $W_{t,T} = \widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)$ I arrive at the following result:

$$\sup_{u\in I_{h_2}} \left| \widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u) - \mathbb{E}\left[\widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)\right] \right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}\right).$$

iii)

$$\sup_{u\in I_{h_2}} \left| \mathbb{E}\left[\widehat{\mathcal{I}}_1(u) - \sigma^2(u)\widehat{f}(u)\right] \right| = O_p\left(h_2^2\right).$$

(iv)

$$\sup_{u\in I_{h_2}} \left| \widehat{\mathcal{I}}_2(u) \right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}\right).$$

(v)

$$\sup_{u\in I_{h_2}} \left| \widehat{\mathcal{I}}_3(u) \right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right).$$

(vi)

$$\sup_{u\in I_{h_2}} \left| \widehat{\mathcal{I}}_4(u) - \mathcal{I}_4(u) \right| = O_p\left(\frac{1}{T^2h_1^2}\sqrt{\frac{\log T}{Th_2}}\right).$$

Proof of (iv). First I re-write $\mathcal{I}_2(u)$ as follows:

$$\widehat{\mathcal{I}}_2(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\left[\sigma^2(t/T) - \sigma^2(u)\right]\{\varepsilon_t^2 - 1\} + \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(u)\{\varepsilon_t^2 - 1\}.$$

Therefore using the fact that $|t/T - u| \leq h^\star := \min\{h_1, h_2\}$:

$$\begin{aligned}
\left|\widehat{\mathcal{I}}_2(u)\right| &= \frac{1}{T}\sum_{t=1}^{T}\left|K_{h_2}\left(t/T - u\right)]\{\varepsilon_t^2 - 1\}\right|\left|\sigma^2(t/T) - \sigma^2(u)\right| + \\
&\qquad + \left|\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(u)\{\varepsilon_t^2 - 1\}\right| \leq \\
&\leq Ch^\star\left|\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)]\{\varepsilon_t^2 - 1\}\right| + \left|\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T - u\right)\sigma^2(u)\{\varepsilon_t^2 - 1\}\right|.
\end{aligned}$$

Applying Theorem 4.1. of Vogt (2012) by setting $d = 0$ and $W_{t,T} = \varepsilon_t^2 - 1$, I arrive at the following result:

$$\sup_{u\in I_{h_2}} \left| \widehat{\mathcal{I}}_2(u) \right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}\right).$$

Proof of (v). I next turn to $\widehat{\mathcal{I}}_3$:

$$\widehat{\mathcal{I}}_3(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\mathbb{Z}_t^T\{\widehat{\theta}(t/T)-\theta(t/T)\}\xi_t =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathbb{Z}_t^T\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u)+\Sigma_T^{-1}(u)\widetilde{W}_T(u)\right] = \widehat{\mathcal{I}}_{31}(u)+\widehat{\mathcal{I}}_{32}(u).$$

In addition, for the simplicity of exposition of further results, I partition $\Sigma(u)^{-1}$ into 4 submatrices each of dimension $(d+1)\times(d+1)$:

$$\Sigma(u)^{-1} = \begin{bmatrix}\widetilde{\Sigma}_{11}(u) & \widetilde{\Sigma}_{12}(u)\\ \\ \widetilde{\Sigma}_{21}(u) & \widetilde{\Sigma}_{22}(u)\end{bmatrix}.$$

In addition I denote $\{\ddot{\rho}(u)\}_{1:(d+1)}$ denotes the first $d+1$ elements of $\ddot{\rho}(u)$ and $\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}$ denotes the last $d+1$ elements of $\ddot{\rho}(u)$. I start with $\widehat{\mathcal{I}}_{31}(u)$:

$$\widehat{\mathcal{I}}_{31}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathbb{Z}_t^T\Sigma_T^{-1}(u)\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathbb{Z}_t^T H^{-1}\left[H^{-1}\Sigma_T(u)H^{-1}\right]^{-1}H^{-1}\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathbb{Z}_t^T H^{-1}\left[\Sigma^{-1}(u)+o_p(1)\right]H^{-1}\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u) =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathrm{tr}\left[\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u)\right] =$$

$$= \frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t \mathrm{tr}\left[\Sigma^{-1}(u)H^{-1}\mathbb{B}_T(u)\mathbb{Z}_t^T H^{-1}\right] = \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t\times$$

$$\times\left\{\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^2\mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)H^{-1}\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T H^{-1}\right]+\right.$$

$$\left.+\sum_{t=1}^{T} K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^4\mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)H^{-1}\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T H^{-1}\right]\right\} =$$

$$= \frac{h_1^2}{T^2 h_1 h_2}\sum_{t=1}^{T}\sum_{s=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t K\left(\frac{s/T-u}{h_1}\right)\times$$

$$\times\left\{\left(\frac{s/T-u}{h_1}\right)^2\mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_s^T\right]+\right.$$

$$\left.+\left(\frac{s/T-u}{h_1}\right)^4\mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_s^T\right]\right\}.$$

Therefore, provided $|K(x)| \leq 1$:

$$\left|\widehat{\mathcal{I}}_{31}(u)\right| \leq \frac{h_1^2}{Th_1h_2}\left|\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t K\left(\frac{s/T-u}{h_1}\right)\times \right.$$
$$\times\left\{\left(\frac{s/T-u}{h_1}\right)^2 \mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_s^T\right]+\right.$$
$$\left.\left.+\left(\frac{s/T-u}{h_1}\right)^4 \mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_s\mathbb{X}_s^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_s^T\right]\right\}\right| \leq$$

$$\leq \frac{h_1^2}{Th_1h_2}\left|\frac{1}{T}\sum_{t=1}^{T} K\left(\frac{t/T-u}{h_2}\right)K\left(\frac{t/T-u}{h_1}\right)\xi_t\left\{\left(\frac{t/T-u}{h_1}\right)^2 \mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right]+\right.\right.$$
$$\left.\left.+\left(\frac{t/T-u}{h_1}\right)^4 \mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T\right]\right\}\right| \leq$$
$$\leq \frac{h_1^2}{Th_1}\left|\frac{1}{Th_2}\sum_{t=1}^{T} K\left(\frac{t/T-u}{h_2}\right)\xi_t\left\{\left(\frac{t/T-u}{h_1}\right)^2 \mathrm{tr}\left[\widetilde{\Sigma}_{11}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right]+\right.\right.$$
$$\left.\left.+\left(\frac{t/T-u}{h_1}\right)^4 \mathrm{tr}\left[\widetilde{\Sigma}_{22}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{(d+2):2(d+1)}\mathbb{X}_t^T\right]\right\}\right|$$

Finally applying Theorem 4.1 in Vogt (2012) by setting $d = 0$ and

$$W_{t,T} = \xi_t\left(\frac{t/T-u}{h_1}\right)^j \mathrm{tr}\left[\widetilde{\Sigma}_{ii}(u)\mathbb{X}_t\mathbb{X}_t^T\{\ddot{\rho}(u)\}_{1:(d+1)}\mathbb{X}_t^T\right], \qquad \text{for} \qquad i=1,2 \quad \text{and} \quad j=2,4.$$

yields the final result:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{31}(u)\right| = O_p\left(\frac{h_1}{T}\sqrt{\frac{\log T}{Th_2}}\right). \tag{3.31}$$

I now consider $\widehat{\mathcal{I}}_{32}(u)$:

$$\widehat{\mathcal{I}}_{32}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_2}(t/T-u)\,\xi_t\mathbb{Z}_t^T H^{-1}\left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}H^{-1}\widetilde{W}_T(u) =$$
$$= \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}(t/T-u)\,\xi_t\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)\sum_{t=1}^{T} K_{h_1}(t/T-u)\,H^{-1}\mathbb{Z}_t\xi_t =$$
$$= \frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}(t/T-u)\,K_{h_1}(t/T-u)\,\xi_t^2\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t +$$
$$+ \frac{2}{T^2}\sum_{j=1}^{T-1} K_{h_2}(t/T-u)\,K_{h_1}((t-j)/T-u)\,\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j}.$$

In addition, it holds that

$$\left|\widehat{\mathcal{I}}_{32}(u)\right| = \left|\frac{1}{T^2}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right) K_{h_1}\left(t/T-u\right)\xi_t^2 \mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t + \right.$$

$$\left. + \frac{2}{T^2}\sum_{j=1}^{T-1} K_{h_2}\left(t/T-u\right) K_{h_1}\left((t-j)/T-u\right)\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j}\right| \leq$$

$$\leq \frac{1}{Th_1}\left|\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t^2\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_t\right| +$$

$$+ \frac{2}{Th_1}\left|\frac{1}{T}\sum_{j=1}^{T-1} K_{h_2}\left(t/T-u\right)\xi_t\xi_{t-j}\mathbb{Z}_t^T H^{-1}\Sigma^{-1}(u)H^{-1}\mathbb{Z}_{t-j}\right|.$$

Therefore applying Theorem 4.1 of Vogt (2012) to the appropriate terms in the above expression I arrive at the following results:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{32} - \mathbb{E}\left[\widehat{\mathcal{I}}_{32}\right]\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right), \tag{3.32}$$

and

$$\sup_{u\in I_{h_2}}\left|\mathbb{E}\left[\widehat{\mathcal{I}}_{32} - \mathcal{I}_{32}\right]\right| = O_p\left(\frac{1}{Th_1}\right) \tag{3.33}$$

and

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_{32}(u) - \mathcal{I}_{32}\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right). \tag{3.34}$$

Combining (3.31)-(3.34) yields:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_3(u) - \mathcal{I}_3(u)\right| = O_p\left(\frac{1}{Th_1}\sqrt{\frac{\log T}{Th_2}}\right).$$

Proof of (vi). Finally I address $\widehat{\mathcal{I}}_4(u)$ term:

$$\widehat{\mathcal{I}}_4(u) =$$

$$= \frac{1}{T}\sum_{t}^{T} K_{h_2}\left(t/T-u\right)\mathbb{Z}_t\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u) + \Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u) + \Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]^T\mathbb{Z}_t =$$

$$= \frac{1}{T}\sum_{t}^{T} K_{h_2}\left(t/T-u\right)\operatorname{tr}\left\{\mathbb{Z}_t\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u) + \Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]\times\right.$$

$$\left.\times\left[\Sigma_T^{-1}(u)\mathbb{B}_T(u) + \Sigma_T^{-1}(u)\widetilde{W}_T(u)\right]^T\mathbb{Z}_t\right\} = \frac{1}{T}\sum_{t}^{T} K_{h_2}\left(t/T-u\right)\operatorname{tr}\left\{\left[\Sigma_T^{-1}(u)\left(\mathbb{B}_T(u)\mathbb{B}_T^T(u) + \right.\right.\right.$$

$$\left.\left.\left. + 2\mathbb{B}_T(u)\widetilde{W}_T^T(u) + \widetilde{W}_T(u)\widetilde{W}_T^T(u)\right)\Sigma_T^{-1}(u)\right]\mathbb{Z}_t\mathbb{Z}_t^T\right\}.$$

Taking the expectation of the above expression, I arrive at

$$\mathbb{E}\left[\widehat{\mathcal{I}}_4(u)\right] = \frac{1}{T}\mathbb{E}\sum_t^T K_{h_2}\left(t/T-u\right)\mathrm{tr}\Bigg\{\Bigg[\Sigma_T^{-1}(u)\Bigg(\mathbb{B}_T(u)\mathbb{B}_T^T(u)+$$

$$+2\mathbb{B}_T(u)\widetilde{W}_T^T(u)+\widetilde{W}_T(u)\widetilde{W}_T^T(u)\Bigg)\Sigma_T^{-1}(u)\Bigg]\mathbb{Z}_t\mathbb{Z}_t^T\Bigg\} =$$

$$= \frac{1}{T}\mathbb{E}\sum_t^T K_{h_2}\left(t/T-u\right)\mathrm{tr}\Bigg\{\Bigg[\left(H^{-1}\Sigma_T(u)H^{-1}\right)^{-1}\Bigg(H^{-1}\mathbb{B}_T(u)\mathbb{B}_T^T(u)H^{-1}+$$

$$+H^{-1}\widetilde{W}_T(u)\widetilde{W}_T^T(u)H^{-1}\Bigg)\left(H^{-1}\Sigma_T(u)H^{-1}\right)\Bigg]\mathbb{Z}_t\mathbb{Z}_t^T\Bigg\} =$$

$$= \frac{1}{T}\mathbb{E}\Bigg\{K_{h_2}\left(t/T-u\right)\mathrm{tr}\Bigg[\Sigma^{-1}(u)\Bigg(H^{-1}\mathbb{B}_T(u)\mathbb{B}_T^T(u)H^{-1}+$$

$$+H^{-1}\widetilde{W}_T(u)\widetilde{W}_T^T(u)H^{-1}\Bigg)\Sigma^{-1}(u)\mathbb{Z}_t\mathbb{Z}_t^T\Bigg]\Bigg\}.$$

Using similar steps as for proving v) I arrive at the following result:

$$\sup_{u\in I_{h_2}}\left|\widehat{\mathcal{I}}_4(u)-\mathcal{I}_4(u)\right| = O_p\left(\frac{1}{T^2h_1^2}\sqrt{\frac{\log T}{Th_2}}\right).$$

And finally combining results from (i)-(vi) yields:

$$\sup_{u\in I_{h_2}}\left|\widehat{\sigma}^2(u)-\sigma^2(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}}+h_2^2\right), \tag{3.35}$$

which completes the proof. ■

Proofs of Theorems 3-6.

In what follows I show the proof of Theorem 6 for the extended statistics $\mathcal{S}_T'$ since the basic statistics $\mathcal{S}_T$ (without an extra weighting $\phi(u)$) is nested in $\mathcal{S}_T'$ and can be obtained by setting $\phi(u)=1\ \ \forall\ u\in[0,1]$. Proofs of Theorems 3 and 4 can be obtained as a special case of the proof of Theorem 6, i.e. by setting the function $\triangle$ to zero in the proofs. Proof of Theorem 5 is straightforward provided the proof of Theorems 3 and therefore is omitted. I start by rewriting the test statistics $\mathcal{S}_T'$, normalised by its rate as follows:

$$\sqrt{T}\mathcal{S}_T' = \frac{\sqrt{T}}{\sqrt{\Phi}}\int_0^1 \phi(u)\widehat{\tau}(u)du = \sqrt{T}\int_0^1\left[V_T(u)+\mathcal{B}_T(u)\right]du,$$

where $\Phi = \int_0^1 \phi^2(u)du$ and with the notation from (3.26) I can write

$$V_T(u) = \frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\widetilde{W}_{T,0}(u)}{\sqrt{\Phi}\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}},$$

and

$$\mathcal{B}_T(u) = \frac{1}{2}\frac{\phi(u)h_1^2X_t^T(u)\Sigma_{T,0}^{-1}(u)\Big(\Sigma_{T,2}(u)\ddot{\rho}(u) + \mathbb{R}_{T,0}(u)\Big)}{\sqrt{\Phi}\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}} + \\ + c_T \triangle(u) + c_T\Big\{\triangle(t/T) - \triangle(u)\Big\},$$

where I used Theorem 2 to substitute $\widehat{\sigma}(u)$ with $\sigma(u)$ and where $\mathbb{R}_{T,0}(u)$ is the residual part, defined in (3.24)-(3.25) and

$$\Sigma_{T,m}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}(t/T - u)(t/T - u)^m \mathbb{X}_t\mathbb{X}_t^T,$$

$$\widetilde{W}_{T,0}(u) = \frac{1}{T}\sum_{t=1}^{T} K_{h_1}(t/T - u)\mathbb{X}_t\xi_t,$$

where again recall that for brevity I write $\mathbb{X}_t$ to abbreviate $\mathbb{X}_{t,T}$. Proof of the Theorem 6 follows from the following two lemmas.

Lemma 1. *Under the assumptions (A1)-(A5), it holds that*

$$\sqrt{T}\int_0^1 V_T(u)du \xrightarrow{d} \mathcal{N}(0,1).$$

Lemma 2. *Under the assumptions (A1)-(A5), it holds that*

$$\sqrt{T}\int_0^1 \mathcal{B}_T(u)du = \int_0^1 \triangle(u)du + h_1^2\sqrt{T}\mathbb{B}_T + o_p(1),$$

where

$$\mathbb{B}_T = \frac{1}{2\sqrt{\Phi}}\int_0^1 \frac{\phi(u)\lambda_2\mathbb{X}_t^T(u)\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}}du.$$

Proof of Lemma 1.

For the ease of exposition I will need to introduce some further notation. First, denote by $\delta(u) = \sigma(u)\sqrt{\nu_0 \mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)}$ and denote by $Y_{t,T}$ the following quantity:

$$Y_{t,T} = \frac{\sqrt{h_1}}{\sqrt{T\Phi}} \int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t du.$$

It is then straightforward to verify that $Y_{t,T}$ is a martingale difference array since conditional on the $X_{t,T}$:

$$\mathbb{E}\left[Y_{t,T}|\mathcal{F}_{t-1,T}, \mathbb{X}_{t,T}\right] = \frac{\sqrt{h_1}}{\sqrt{T\Phi}} \int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{E}\left[\mathbb{X}_t\xi_t|\mathcal{F}_{t-1,T}\right]du + o(1) = 0,$$

where $\mathcal{F}_{t-1,T} := \sigma\left(\mathbb{X}_{s,T}, \xi_{s,T} : s \leq t-1\right)$ denotes the sigma-algebra induced by the history of $\mathbb{X}_{t,T}$ and $\xi_{t,T}$. I can therefore apply the central limit theorem for the martingale difference arrays (e.g. Theorem 3.2 in Hall and Heyde (1980)) to establish that $\sum_{t=1}^{T} Y_{t,T}$ is asymptotically normal. For applying Theorem 3.2 in Hall and Heyde (1980), it suffices to verify the following conditions:

(C1) $\sum_{t=1}^{T}\mathbb{E}\left[Y_{t,T}^2|\mathcal{F}_{t-1,T}, \mathbb{X}_{t,T}\right] \xrightarrow{p} V_1$,

(C2) for every $\epsilon > 0$, it holds that $\sum_t \mathbb{E}\left[Y_{t,T}^2\{|Y_{t,T} > \epsilon|\}|\mathcal{F}_{t-1,T}, \mathbb{X}_{t,T}\right] \xrightarrow{p} 0$.

Proof of (C1).

I first calculate $Y_{t,T}^2$, which is given by the following expression:

$$Y_{t,T}^2 = \frac{h_1}{T\Phi}\left(\int_0^1 \delta^{-1}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t\right)^2 =$$

$$= \frac{h_1}{T\Phi}\int_0^1 \delta^{-2}(u)\phi^2(u)K_{h_1}^2\left(t/T-u\right)\xi_t^2\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u)\mathbb{X}_t(u)du +$$

$$+\frac{h_1}{T\Phi}\int_0^1\int_0^1 \delta^{-2}(u)\phi^2(u)K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\xi_t^2\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u')\mathbb{X}_t(u')dudu' + o_p(1).$$

Taking expectation of the above expression

$$\sum_{t=1}^{T}\mathbb{E}\left[Y_{t,T}^{2}|X_t(u)\right]=$$

$$=\frac{h_1}{T\Phi}\sum_{t=1}^{T}\int_0^1\delta^{-2}(u)\phi^2(u)K_{h_1}^2\left(t/T-u\right)\sigma^2(t/T)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{E}\left[\varepsilon_t^2\mathbb{X}_t\mathbb{X}_t^T\right]\Omega^{-1}(u)\mathbb{X}_t(u)du+$$

$$+\frac{h_1}{T\Phi}\sum_0^1\int_0^1\int_0^1\delta^{-2}(u)\phi^2(u)K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\sigma^2(t/T)\mathbb{E}\left[\mathbb{X}_t^T(u)\varepsilon_t^2\mathbb{X}_t\mathbb{X}_t^T\Omega^{-1}(u')\mathbb{X}_t(u')\right]dudu'=$$

$$==\frac{1}{h_1\Phi}\int_0^1\int_0^1\delta^{-2}(u)\phi^2(u)K^2\left(\frac{y-u}{h_1}\right)\sigma^2(y)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)dudy=$$

$$=\frac{1}{\Phi}\int_0^1\delta^{-2}(u)\phi^2(u)\nu_0\sigma^2(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t(u)du=\frac{1}{\Psi}\int_0^1\phi^2(u)du=1,$$

where I used the fact that $\delta^{-2}(u)=\sigma^2(u)\nu_0\mathbb{X}_t^T(u)\Omega^{-1}(u)\mathbb{X}_t$. I now establish the following intermediate result. I calculate the covariance between $\sqrt{Th_1}\widetilde{W}_{T,0}(u)$ and $\sqrt{Th_1}\widetilde{W}_{T,0}(u')$ for generic rescaled time points $u,u'\in[0,1]$. Without loss of generality, assume that $u'<u$ for $(u-h_1)T\leq t\leq(u+h_1)T$ and $(u'-h_1)T\leq t\leq(u'+h_1)T$, then it holds that

$$\mathbb{E}\left(\frac{h_1}{T}\sum_{t=1}^{T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=1}^{T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$=\mathbb{E}\left(\frac{h_1}{T}\sum_{t=(u-h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$=\mathbb{E}\Bigg\{\frac{h_1}{T}\Bigg[\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t+\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\Bigg]\times$$

$$\times\Bigg[\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\Bigg]\Bigg\}=$$

$$=\mathbb{E}\Bigg\{\frac{h_1}{T}\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+$$

$$+\frac{h_1}{T}\sum_{t=(u-h)T}^{(u'+h)T-1}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+$$

$$+\frac{h_1}{T}\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u'-h)T}^{(u-h)T-1}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t+$$

$$+\frac{h_1}{T}\sum_{t=(u'+h)T}^{(u+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\Bigg\}=$$

$$=\frac{h_1}{T}\mathbb{E}\left(\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)\mathbb{X}_t(u)\sigma(t/T)\varepsilon_t\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u'\right)\mathbb{X}_t^T(u')\sigma(t/T)\varepsilon_t\right)=$$

$$\frac{h_1}{T}\mathbb{E}\Bigg\{\sum_{t=(u-h)T}^{(u'+h)T}K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_t^T(u')\sigma^2(t/T)\varepsilon_t^2+$$

$$+\underbrace{\sum_{t=(u-h)T}^{(u'+h)T}\sum_{s=(u-h)T}^{(u'+h)T}}_{t\neq s}K_{h_1}\left(t/T-u\right)K_{h_1}\left(s/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_s^T(u')\sigma(t/T)\sigma(s/T)\varepsilon_t\varepsilon_s\Bigg\}=$$

$$=h_1\mathbb{E}\left(K_{h_1}\left(t/T-u\right)K_{h_1}\left(t/T-u'\right)\mathbb{X}_t(u)\mathbb{X}_t^T(u')\sigma^2(t/T)\varepsilon_t^2\right)=$$

$$=\frac{h_1}{T}\sum_{t=1}^{T}K_{h_1}(t/T-u)K_{h_1}(t/T-u')\Omega(t/T)\sigma^2(t/T)+o(1)=$$

$$=\int_0^1 K\left(y\right)K\left(y+\frac{u-u'}{h_1}\right)y\sigma^2(yh_1+u)\Omega(yh_1+u)dy=\nu_0\sigma^2(u)\Omega(u)+o(1).\qquad\blacksquare$$

Proof of (C2).

Provided that $\mathbb{E}|\mathbb{X}_{it}|^{2+\delta}<\infty\ \forall\ i=1,2,\cdots,d+1$ and $\forall\ t$ and $\mathbb{E}|\varepsilon_{it}|^{2+\delta}<\infty$, then for any large $C<\infty$ I can further write the following:

$$\sum_{t=1}^{T}\mathbb{E}\left[(Y_{t,T})^2\{|Y_{t,T}|>\epsilon\}|\mathcal{F}_{t-1,T}\right]\leq\sum_{t}\mathbb{E}\left[\frac{|Y_{t,T}|^{2+\delta}}{|Y_{t,T}|^{\delta}}\{|Y_{t,T}|>\epsilon\}|\mathcal{F}_{t-1,T}\right]=$$

$$=\sum_{t}\mathbb{E}\left[|Y_{t,T}|^{2+\delta}\{|Y_{t,T}|^{1/\delta}>\epsilon\}|\mathcal{F}_{t-1,T}\right]\leq\frac{1}{T^{\delta/2}\epsilon^{\delta}}\mathbb{E}\left[|Y_{t,T}|^{2+\delta}\right]\leq\frac{C}{T^{\delta/2}\epsilon^{\delta}}\xrightarrow{p}0.$$

Combining all of the above derivations proves Lemma 1. ■

Proof of Lemma 2. First recall from (3.27) that

$$h_1^{-m}\Sigma_{T,m}=\lambda_m\Omega(u)\{1+o_p(1)\}\qquad\text{and}\qquad h_1^{-m}\mathbb{R}_{T,m}=o_p(h_1^2).$$

With the notation introduced in the proof of Lemma 1, we can write

$$\mathcal{B}_T(u)=\frac{1}{2\sqrt{\Phi}}h_1^2\phi(u)\delta^{-1}(u)X_t^T(u)\ddot{\rho}(u)\frac{1}{T}\sum_{t=1}^{T}K_{h_1}\left(t/T-u\right)\left(t/T-u\right)^2\mathbb{X}_t\mathbb{X}_t^T+c_T\triangle(u)+o_p(h_1^2).$$

Setting $c_T=1/\sqrt{T}$, normalising $\mathcal{B}_T(u)$ by $\sqrt{T}$ and taking expectation yields:

$$\mathbb{E}\left[\sqrt{T}\int_0^1\mathcal{B}_T(u)du\right]=\frac{\sqrt{T}h_1^2}{2\sqrt{\Phi}}\int_0^1\frac{\phi(u)X_t^T(u)\lambda_2\ddot{\rho}(u)}{\sigma(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du+\int_0^1\triangle(u)du+o(1),$$

which concludes the proof of Lemma 2. ■

Proof of Theorem 7.

I start by deriving $Pr\left(\Delta\mathcal{L}_{T+1}\leq 0\right)$ using my model (3.12):

$$Pr\left(\Delta\mathcal{L}_{T+1}\leq 0|\mathcal{F}_T\right)=Pr\left(\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)+\sigma\left(\frac{T+1}{T}\right)\varepsilon_{T+1}\leq 0|\mathcal{F}_T\right)=$$

$$=Pr\left(\varepsilon_{T+1}\leq\frac{-\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right).$$

Moreover, provided the assumption that

$$\rho\left(\frac{T+1}{T}\right)\approx\rho\left(\frac{T}{T}\right)=\rho(1)\qquad\text{and}\qquad\sigma\left(\frac{T+1}{T}\right)\approx\sigma\left(\frac{T}{T}\right)=\sigma(1).$$

It then holds that

$$Pr\left(\varepsilon_{T+1}\leq\frac{-\mathbb{X}_{T+1}^T\rho\left(\frac{T+1}{T}\right)}{\sigma\left(\frac{T+1}{T}\right)}\right)\approx Pr\left(\varepsilon_{T+1}\leq\frac{-\mathbb{X}_{T+1}^T\rho\left(1\right)}{\sigma\left(1\right)}\right)=:F\left(\varepsilon^{\star}(1)\right),$$

where $\varepsilon^\star(1) := \frac{-\mathbb{X}_{T+1}^T\rho(1)}{\sigma(1)}$. Now, conditional on the sample $\{\Delta\mathcal{L}_t\}_{t=1}^T$:

$$\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = \widehat{F}\left(\widehat{\varepsilon}^\star(1)\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\left(\widehat{\varepsilon}_t \leq \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right), \tag{3.36}$$

where $\widehat{\varepsilon}^\star(1) := \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}(1)}{\widehat{\sigma}(1)}$. Taking expectation of the above and using a first-order Taylor expansion it holds that:

$$\begin{aligned}\mathbb{E}\left[\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right)\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\widehat{\varepsilon}_t \leq \widehat{\varepsilon}^\star(1)\middle|\widehat{\varepsilon}^\star\right)\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}\left(\widehat{\varepsilon}_t \leq \widehat{\varepsilon}^\star(1)\middle|\widehat{\varepsilon}^\star\right)\right]\right] = \\ = \mathbb{E}\left[F\left(\varepsilon^\star(1)\right) + F'\left(\varepsilon^\star(1)\right)\left(\widehat{\varepsilon}^\star(1) - \varepsilon^\star(1)\right) + o(T^{-1/2})\right] = \\ = F(\varepsilon^\star(1)) + f(\varepsilon^\star(1))\mathbb{E}\left\{\left(\widehat{\varepsilon}^\star(1) - \varepsilon^\star(1)\right)\right\} + o(T^{-1/2}).\end{aligned}$$

Using a first-order Taylor expansion of $\widehat{\varepsilon}^\star$ around $\varepsilon^\star$, it holds that:

$$\begin{aligned}\widehat{\varepsilon}^\star(1) = \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}\left(1\right)}{\widehat{\sigma}(1)} = \frac{-\mathbb{X}_{T+1}^T\rho\left(1\right)}{\sigma(1)} - \frac{\mathbb{X}_{T+1}^T\left[\widehat{\rho}(1) - \rho\left(1\right)\right]}{\widehat{\sigma}(1)} + \mathbb{X}_{T+1}^T\frac{\rho\left(1\right)\left[\widehat{\sigma}(1) - \sigma(1)\right]}{\widehat{\sigma}^2(1)} = \\ = \varepsilon^\star(1) - \frac{\mathbb{X}_{T+1}^T\left[\widehat{\rho}(1) - \rho\left(1\right)\right]}{\sigma(1)} + \frac{\mathbb{X}_{T+1}^T\rho\left(1\right)\left[\widehat{\sigma}(1) - \sigma(1)\right]}{\sigma^2(1)},\end{aligned}$$

and therefore

$$\mathbb{E}\left[\widehat{\varepsilon}^\star(1) - \varepsilon(1)\right] = \frac{1}{2\sigma^2(1)}\mathbb{X}_{T+1}(1)\left\{h_1^2\lambda_2\ddot{\rho}(1)\sigma(1) + h_2^2\lambda_2\ddot{\sigma}(1)\right\} =: \mathbb{B}_3(1).$$

Given that $\mathbb{1}\left(\widehat{\varepsilon}_t \leq \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}(1)}{\widehat{\sigma}(1)}\right)$ is a Bernulli random variable and therefore

$$\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}\left(\widehat{\varepsilon}_t \leq \frac{-\mathbb{X}_{T+1}^T\widehat{\rho}\left(1\right)}{\widehat{\sigma}\left(1\right)}\right)$$

has a binomial distribution which as $T \to \infty$ becomes a normal distribution with the mean and variance given by

$$\begin{aligned}\text{var}\left(\sqrt{T}\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right)\right) = \mathbb{E}\left[T\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right)^2\right] - \left(T\mathbb{E}\left[\widehat{F}\left(\widehat{\varepsilon}^\star(1)\right)\right]\right)^2 = \\ = F(\varepsilon^\star(1))\left(1 - F\left(\varepsilon^\star(1)\right)\right) + o(1).\end{aligned}$$

Therefore, it holds that

$$\sqrt{T}\left[\widehat{F}(\widehat{\varepsilon}^\star(1)) - F(\varepsilon^\star(1)) - \mathbb{B}_3(1)\right] \xrightarrow{d} \mathcal{N}\left(0, F\left(\varepsilon^\star(1)\right)\left(1 - F\left(\varepsilon^\star(1)\right)\right)\right), \tag{3.37}$$

where

$$\mathbb{B}_3(1) = \frac{f\left(\varepsilon^\star(1)\right)}{2\sigma^2(1)}\mathbb{X}_{T+1}(1)\Big\{h_1^2\lambda_2\ddot{\rho}(1)\sigma(1) + h_2^2\lambda_2\ddot{\sigma}(1)\Big\},$$

which completes the proof. ■

Proof of Theorem 8.

The proof of the Theorem 8 closely follows original proof of Theorems 3-6 with the bootstrapped quantities, denoted by $\star$. In particular, $\mathbb{E}^\star(\cdot)$, $\text{var}^\star(\cdot)$ and $\mathbb{P}^\star(\cdot) := \mathbb{P}^\star\left(\cdot|\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T\right)$ are used to denote the expectation, variance and the distribution respectively conditional on the sample $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}_{t=1}^T$. I start by making use of the following notation:

$$\sqrt{T}\mathcal{S}_T^\star = \frac{\sqrt{T}}{\sqrt{\Phi}}\int_0^1 \phi(u)\widehat{\tau}^\star(u)du = \sqrt{T}\int_0^1 \left[V_T^\star(u) + \mathcal{B}_T^\star(u)\right]du,$$

where $\Phi = \int_0^1 \phi^2(u)du$ and with the notation from (3.26) I can write

$$V_T^\star(u) = \frac{\sqrt{Th_1}\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\widetilde{W}_{T,0}^\star(u)}{\sqrt{\Phi}\widehat{\sigma}^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}, \tag{3.38}$$

and

$$\mathcal{B}_T^\star(u) = \frac{\phi(u)}{\sqrt{\Phi}}\tau^\star(u) + \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\widetilde{\mathcal{R}}_T^\star(u) + o_p(h_1^2)\right)}{\sqrt{\Phi}\widehat{\sigma}^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}, \tag{3.39}$$

where I used Theorem 2 to substitute $\widehat{\sigma}^\star(u)$ with $\sigma^\star(u)$ and where the following notation is used:

$$\Sigma_{T,m}(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T - u\right)\left(t/T - u\right)^m \mathbb{X}_t(u)\mathbb{X}_t^T(u),$$

$$\widetilde{W}_{T,0}^\star(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T - u\right)\mathbb{X}_t(u)\xi_t^\star,$$

and

$$\widetilde{\mathcal{R}}_T^\star(u) = \frac{1}{T}\sum_{t=1}^T K_{h_1}\left(t/T - u\right)\mathbb{X}_t(u)\mathbb{X}_t^T\Big\{\rho(t/T) - \widetilde{\rho}(u)\Big\}.$$

In what follows I need to show that under the conditions of Theorem 8, the following holds:

$$\sqrt{T}\int_0^1 V_T^\star(u)du \xrightarrow{d} \mathcal{N}\left(0,1\right), \tag{B1}$$

conditional on the sample $\{\Delta\mathcal{L}_{t,T}, \mathbb{X}_{t,T}\}$ with probability tending to one, and

$$\sqrt{T}\int_0^1 \mathcal{B}_T^\star(u)du = h_1^2\sqrt{T}\mathbb{B}_T + o_p(1). \tag{B2}$$

For the proofs of (B1) and (B2) I will be using the following notation from the proof of Lemma 1: denote by $\delta^\star(u) = \sigma^\star(u)\sqrt{\nu_0\mathbb{X}_t(u)\Omega^{-1}(u)\mathbb{X}_t(u)}$ and

$$Y_{t,T}^\star = \frac{\sqrt{h_1}}{\sqrt{T\Phi}}\int_0^1 \delta^{-1,\star}(u)\phi(u)\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)K_{h_1}\left(t/T-u\right)\mathbb{X}_t\xi_t^\star du, \tag{3.40}$$

where $\xi_t^\star = \widehat{\xi}_t\eta_t$ are the bootstrapped residuals. Note that since η_t are i.i.d. it then follows that $\xi_t^\star$ have the same mixing properties as the original residuals ξ_t (see Theorem 5.2 in Bradley (2005)). It is than straightforward to establish that $Y_{t,T}^\star$ is also a martingale difference sequence and by using uniform convergence results in Theorem 2, in what follows I establish that conditional on the sample with probability one it holds that

$$\sqrt{T}\int_0^1 \mathcal{B}_T^\star(u)du = \sqrt{T}\int_0^1 \mathcal{B}_T(u)du + o_p(1), \tag{B3}$$

and

$$\mathbb{P}^\star\left(\sqrt{T}\int_0^1 V_T^\star(u)du \le x\right) \xrightarrow{p} \Phi(x), \tag{B4}$$

where $\Phi(x)$ is the standard Gaussian distribution. Therefore,

$$\mathbb{P}^\star\left(\sqrt{T}\left(S_T^\star - \mathbb{B}_T\right) \le x\right) \xrightarrow{p} \Phi(x),$$

which then completes the proof of Theorem 8. ■

Below I prove (B3) and (B4). However, before proving (B3) and (B4), I first consider $\widehat{\sigma}^\star(u)$. I can write

$$\widehat{\sigma}^{\star}(u) = \frac{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)(\widehat{\xi}_t^{\star})^2}{\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)} = \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)(\widehat{\xi}_t^{\star})^2 =$$

$$= \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\Delta\mathcal{L}_t^{\star} - \mathbb{X}_t^T\widehat{\rho}^{\star}(t/T)\right)^2 =$$

$$= \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\left(\mathbb{X}_t^T\widetilde{\rho}(t/T) + \xi_t^{\star} - \mathbb{X}_t^T\widehat{\rho}^{\star}(t/T)\right)^2 = \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)(\xi_t^{\star})^2 +$$

$$+ \widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\mathbb{X}_t^T\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right)\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right)\mathbb{X}_t +$$

$$+ 2\widehat{f}(u)^{-1}\frac{1}{T}\sum_{t=1}^{T} K_{h_2}\left(t/T-u\right)\xi_t^{\star}\left(\widetilde{\rho}(t/T) - \widehat{\rho}^{\star}(t/T)\right).$$

Using results (i)-(vi) in the proof of Theorem 2, and the definition of $\widehat{\rho}^{\star}(t/T)$, it is then straightforward to establish that

$$\sup_{u\in I_{h_2}}\left|\widehat{\sigma}^{\star}(u) - \sigma^{\star}(u)\right| = O_p\left(\sqrt{\frac{\log T}{Th_2}} + h_2^2\right), \tag{3.41}$$

where I_{h_2} is defined in the statement of Theorem 2. I now prove (B3) and (B4). Using the result in (3.41) I substitute $\widehat{\sigma}^{\star}(u)$ with $\sigma^{\star}(u)$ in the expressions of (3.38), (3.39) and (3.40).

Proof of B3.

It then holds that

$$\sqrt{T}\int_0^1 \mathcal{B}_T^{\star}(u)du = \frac{\sqrt{T}}{\sqrt{\Phi}}\int_0^1 \phi(u)\tau^{\star}(u)du + \sqrt{T}\int_0^1 \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\widetilde{\mathcal{R}}_T^{\star}(u) + o_p(h_1^2)\right)}{\sqrt{\Phi}\widehat{\sigma}^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du =$$

$$= \frac{\sqrt{T}}{2}\int_0^1 \frac{\sqrt{Th_1}\phi(u)h_1^2\mathbb{X}_t^T(u)\Sigma_{T,0}^{-1}(u)\left(\Sigma_{T,2}(u)\ddot{\rho}(u) + o_p(h_1^2)\right)}{\sqrt{\Phi}\sigma^{\star}(u)\sqrt{\nu_0\mathbb{X}_t(u)^T\Omega^{-1}(u)\mathbb{X}_t(u)}}du =$$

$$= \sqrt{T}\int_0^1 \mathbb{B}_T(u)du + o_p(1) \quad , \tag{3.42}$$

where in the second line of (3.42) I used the fact that $\int_0^1 \phi(u)\tau^{\star}(u)du = 0$ by construction in (3.17)-(3.18). This completes the proof of (B3). ■

Proof of B4.

Denoting by $\zeta_t := \xi_t \eta_t$, it holds that

$$\xi_t^\star := \widehat{\xi}_t \eta_t = (\Delta \mathcal{L}_t - \widehat{\mu}_t)\, \eta_t = (\mu_t + \xi_t - \widehat{\mu}_t)\, \eta_t =$$
$$= \xi_t \eta_t + \mathbb{X}_t^T \left(\widehat{\rho}(t/T) - \rho(t/T)\right)\eta_t = \zeta_t + \mathbb{X}_t^T \left(\widehat{\rho}(t/T) - \rho(t/T)\right)\eta_t.$$

First recall that $\widehat{\rho}(u) - \rho(u) = O_p\left(\sqrt{\frac{\log T}{T h_1}} + h_1^2\right)$, and the distribution of $Y_{t,T}^\star$ in (3.40) will be given by the expression involving ζ_t, i.e.

$$\widetilde{Y}_{t,T}^\star = \frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\int_0^1 K_{h_1}\left(t/T - u\right)\mathbb{X}_t(u)\zeta_t du}{\sqrt{\Phi}\sigma^\star(u)\sqrt{\nu_0 \mathbb{X}_t(u)^T \Omega^{-1}(u)\mathbb{X}_t(u)}} =$$
$$= \frac{\sqrt{h_1}\phi(u)\mathbb{X}_t^T(u)\Omega^{-1}(u)\int_0^1 K_{h_1}\left(t/T - u\right)\mathbb{X}_t(u)\xi_t^\star du}{\sqrt{\Phi}\sigma^\star(u)\sqrt{\nu_0 \mathbb{X}_t(u)^T \Omega^{-1}(u)\mathbb{X}_t(u)}},$$

where $\zeta_t := \xi_t \eta_t$. Note also, that provided the definition of η_t, the mixing properties of the original residual sequence ξ_t (Theorem 5.2 of Bradley (2005)) are preserved by the new residual ζ_t and therefore to establish (B4) I need to verify conditions (C1) and (C2). Since the proof follows exactly the same steps as the one in the proof of Theorems 3-6, it is omitted. ■

3.11 Appendix C.

This Appendix presents more results from the application section. In particular, I report results for two more forecast horizons: $k = 6$ months and $k = 12$ months.

Table 3.16: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 6$ months.

		Personal Income					Industrial Production					Producer Price Index					Consumer Price Index				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.49					−1.93					−2.75					−2.25				
	p	0.930					0.980					0.996					0.988				
DI	$\mathcal{S}_T$	1.09	1.54				−0.43	1.82				−3.05	2.78				−3.91	2.39			
	p	0.154	0.068				0.692	**0.040**				0.998	**0.006**				1.00	**0.014**			
AR	$\mathcal{S}_T$	1.42	1.60	1.20			0.82	1.94	1.59			−2.66	2.79	2.71			−4.016	1.82	1.48		
	p	0.084	**0.046**	0.220			0.192	**0.028**	0.056			0.994	**0.008**	**0.008**			1.00	**0.046**	0.062		
RW	$\mathcal{S}_T$	1.46	1.53	0.02	−0.78	–	0.15	1.81	0.82	−0.50	–	3.84	2.55	3.32	6.83	–	2.19	2.31	4.77	7.11	–
	p	0.074	0.062	0.968	0.786	–	0.444	**0.034**	0.200	0.688	–	**0.000**	**0.010**	**0.000**	**0.000**	–	**0.012**	**0.012**	**0.000**	**0.000**	–

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (3.2), for horizon $k = 6$ months. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.49$ (indicating that Lasso is better) with the p-value of 0.930. The p-values in bold indicate rejection of the null (3.2) at the 5% level of significance.

Table 3.17: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 6$ months.

		Personal Income					Industrial Production					Producer Price Index					Consumer Price Index				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.50					−1.87					−2.77					−2.16				
	p	0.140					0.072					**0.016**					**0.040**				
DI	$\mathcal{S}_T$	1.12	1.56				−0.43	1.82				−3.34	2.88				−3.87	2.21			
	p	0.252	0.120				0.640	0.080				**0.004**	**0.004**				**0.000**	**0.030**			
AR	$\mathcal{S}_T$	1.44	1.48	1.21			0.81	1.81	1.60			−2.95	2.92	2.68			−3.86	2.11	1.47		
	p	0.136	0.148	0.240			0.445	0.080	0.108			**0.000**	**0.008**	**0.008**			**0.000**	**0.040**	0.148		
RW	$\mathcal{S}_T$	1.38	1.53	0.02	−0.80	–	0.15	1.82	0.84	−0.51	–	3.83	2.89	3.29	7.11	–	2.21	2.30	4.79	7.12	–
	p	0.188	0.156	0.948	0.456	–	0.780	0.064	0.396	0.640	–	**0.000**	**0.008**	**0.004**	**0.000**	–	**0.036**	**0.032**	**0.000**	**0.000**	–

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds the null of equal predictive ability, see eq. (3.1), for horizon $k = 6$ months. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.50$ (indicating that Lasso is better) with the p-value of 0.140. The p-values in bold indicate rejection of the null (3.1) at the 5% level of significance.

Table 3.18: Results for one-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 12$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.78					−2.06					−1.78					−2.07				
	p	0.960					0.972					0.962					0.980				
DI	$\mathcal{S}_T$	0.90	1.79				−0.66	2.14				−3.42	1.77				−2.80	1.75			
	p	0.192	**0.036**				0.766	**0.028**				0.996	**0.046**				0.998	**0.046**			
AR	$\mathcal{S}_T$	1.82	1.79	0.97			0.79	2.08	1.72			−3.21	1.68	3.52			−4.55	1.93	3.58		
	p	**0.040**	**0.030**	0.178			0.206	**0.020**	**0.034**			1.00	0.060	**0.000**			1.00	**0.034**	**0.000**		
RW	$\mathcal{S}_T$	1.42	1.70	0.34	−0.87	–	0.36	2.14	1.34	−0.34	–	1.51	1.74	6.30	7.21	–	1.36	1.95	3.92	6.15	–
	p	0.076	**0.036**	0.398	0.814	–	0.370	**0.018**	0.092	0.624	–	0.074	**0.048**	**0.000**	**0.000**	–	0.092	**0.032**	**0.000**	**0.000**	–

Note: Table reports the value of the one-sided test statistic $\mathcal{S}_T$ that corresponds to the null of superior predictive ability, see eq. (3.2), for horizon $k = 12$ months. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.78$ (indicating that Lasso is better) with the p-value of 0.960. The p-values in bold indicate rejection of the null (3.2) at the 5% level of significance.

Table 3.19: Results for two-sided test statistic $\mathcal{S}_T$ at nominal size $\alpha = 5\%$ for $k = 12$ months.

		Personal Income					Industrial Production					Producer Price Index					Consumer Price Index				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\mathcal{S}_T$	−1.86					−2.06					−1.75					−1.95				
	p	0.052					**0.036**					0.072					0.060				
DI	$\mathcal{S}_T$	0.89	1.81				−0.69	2.06				−3.53	1.47				−2.80	1.68			
	p	0.380	0.072				0.476	0.052				**0.000**	0.144				**0.008**	0.100			
AR	$\mathcal{S}_T$	1.73	1.76	0.99			0.82	1.89	1.61			−3.29	1.66	3.47			−4.72	1.96	3.59		
	p	0.088	0.092	0.312			0.396	0.084	0.100			**0.000**	0.112	**0.000**			**0.000**	**0.044**	**0.000**		
RW	$\mathcal{S}_T$	1.30	1.80	0.32	−0.87	–	0.36	2.19	1.32	−0.34	–	1.47	1.55	6.42	7.36	–	1.42	1.95	3.74	5.97	–
	p	0.216	0.068	0.772	0.352	–	0.736	**0.032**	0.172	0.768	–	0.144	0.132	**0.000**	**0.000**	–	0.156	**0.050**	**0.000**	**0.000**	–

Note: Table reports the value of the two-sided test statistic $\mathcal{S}_T$ that corresponds the null of superior predictive ability, see eq. (3.1), for horizon $k = 12$ months. The p-values are obtained via the wild bootstrap procedure described in section 3.5. The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$ for which the test statistics $S_T = -1.86$ (indicating that Lasso is better) with the p-value of 0.052. The p-values in bold indicate rejection of the null (3.1) at the 5% level of significance.

Table 3.20: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 6$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.986					0.986					0.999					0.991				
	$\widehat{FI}_l$	0.983					0.979					0.998					0.990				
	$\widehat{FI}_u$	0.988					0.989					1.000					0.992				
	$\widehat{C}$	0.087					0.049					0.028					0.061				
DI	$\widehat{Pr}_{T+1}$	0.542	0.005				0.896	0.031				0.869	0.000				0.852	0.004			
	$\widehat{FI}_l$	0.500	0.003				0.864	0.027				0.831	0.000				0.819	0.003			
	$\widehat{FI}_u$	0.581	0.006				0.906	0.042				0.901	0.001				0.885	0.005			
	$\widehat{C}$	-0.024	-0.082				0.022	-0.072				0.006	-0.028				0.011	-0.050			
AR	$\widehat{Pr}_{T+1}$	0.444	0.005	0.382			0.801	0.017	0.136			0.746	0.001	0.246			0.577	0.000	0.278		
	$\widehat{FI}_l$	0.408	0.003	0.352			0.751	0.013	0.126			0.697	0.000	0.201			0.539	0.000	0.245		
	$\widehat{FI}_u$	0.481	0.007	0.415			0.836	0.025	0.176			0.769	0.002	0.272			0.601	0.001	0.307		
	$\widehat{C}$	-0.040	-0.092	-0.015			-0.012	-0.075	-0.025			0.049	-0.034	0.025			0.039	-0.060	0.016		
RW	$\widehat{Pr}_{T+1}$	0.456	0.008	0.202	0.571		0.935	0.030	0.639	0.883		0.151	0.000	0.088	0.248		0.158	0.000	0.102	0.338	
	$\widehat{FI}_l$	0.424	0.006	0.178	0.537	–	0.916	0.024	0.614	0.859	–	0.127	0.000	0.059	0.231	–	0.147	0.000	0.074	0.329	–
	$\widehat{FI}_u$	0.491	0.011	0.254	0.600	–	0.945	0.036	0.684	0.894	–	0.202	0.001	0.125	0.279	–	0.191	0.001	0.143	0.376	–
	$\widehat{C}$	-0.034	-0.104	-0.026	0.022		-0.031	-0.068	-0.035	-0.001		-0.028	-0.027	-0.012	-0.057		-0.013	-0.052	-0.001	-0.054	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 6$ months. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(3.15). The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.986$ with the corresponding forecast interval $[0.983,\ 0.988]$.

Table 3.21: Sign forecasting for $\Delta\mathcal{L}_{T+1}$ for horizon $k = 12$ months.

		Personal Income					**Industrial Production**					**Producer Price Index**					**Consumer Price Index**				
Benchmark		Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW	Lasso	Bay	DI	AR	RW
Bay	$\widehat{Pr}_{T+1}$	0.415					1.000					0.706					0.821				
	$\widehat{FI}_l$	0.387					0.999					0.683					0.795				
	$\widehat{FI}_u$	0.457					1.000					0.724					0.839				
	$\widehat{C}$	0.072					0.049					0.050					0.102				
DI	$\widehat{Pr}_{T+1}$	0.485	0.540				0.686	0.000				0.823	0.284				0.518	0.197			
	$\widehat{FI}_l$	0.453	0.504				0.659	0.000				0.789	0.262				0.487	0.179			
	$\widehat{FI}_u$	0.526	0.573				0.702	0.001				0.862	0.312				0.561	0.228			
	$\widehat{C}$	-0.017	-0.064				-0.035	-0.054				0.016	-0.073				0.023	-0.101			
AR	$\widehat{Pr}_{T+1}$	0.524	0.560	0.602			0.644	0.000	0.244			0.828	0.326	0.700			0.799	0.197	0.796		
	$\widehat{FI}_l$	0.491	0.523	0.579			0.599	0.000	0.216			0.779	0.305	0.654			0.760	0.176	0.751		
	$\widehat{FI}_u$	0.552	0.588	0.620			0.671	0.001	0.273			0.854	0.352	0.732			0.826	0.230	0.828		
	$\widehat{C}$	-0.018	-0.068	-0.029			-0.011	-0.069	-0.030			0.049	-0.004	0.002			0.031	-0.102	-0.003		
RW	$\widehat{Pr}_{T+1}$	0.491	0.516	0.603	0.372		0.752	0.000	0.734	0.880		0.222	0.276	0.191	0.137		0.234	0.174	0.568	0.110	
	$\widehat{FI}_l$	0.460	0.476	0.573	0.348	–	0.724	0.000	0.702	0.833	–	0.202	0.267	0.154	0.111	–	0.214	0.158	0.532	0.088	–
	$\widehat{FI}_u$	0.520	0.548	0.619	0.394	–	0.770	0.001	0.749	0.897	–	0.268	0.300	0.226	0.193	–	0.276	0.203	0.610	0.157	–
	$\widehat{C}$	-0.026	-0.065	-0.018	0.033		-0.029	-0.072	-0.041	-0.020		-0.017	-0.046	-0.038	-0.057		0.002	-0.097	-0.018	-0.044	

Note: Table reports the results of the sign forecasting for $\Delta\mathcal{L}_{T+1}$ for the forecast horizon $k = 12$ months. $\widehat{Pr}_{T+1}$ is an abbreviation of $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right)$, i.e. the forecasted probability at the very end of the sample. $\widehat{FI}_u$ and $\widehat{FI}_l$ denotes the upper and lower bounds of the forecast interval, such that $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) \in [\widehat{FI}_l,\ \widehat{FI}_u]$. Finally, $\widehat{C}$ denotes the value of the criterion in eq.(3.15). The difference in losses is constructed as the difference between the loss for the column model minus the loss for the row model. For example, $\Delta\mathcal{L}_t^{\text{Lasso,Bay}} = \mathcal{L}_t^{\text{Lasso}} - \mathcal{L}_t^{\text{Bay}}$, for which $\widehat{Pr}\left(\Delta\mathcal{L}_{T+1} \leq 0\right) = 0.415$ with the corresponding forecast interval $[0.387,\ 0.457]$.